ANALOGS: A Computer Program for the Design of Multivariate Sets of Analog Compounds

Mario Marsili*,† and Heinz Saller‡

Department of Chemistry and Materials Science, University of L'Aquila, 67100 L'Aquila, Italy, and
CHEMODATA, Computer-Chemie GmbH and Company KG, 8000 Munich, FRG

Received September 10, 1992

The principles of abstract experimental design are used in a program to generate multivariate sets of chemical substructures as variable substituents of analogs of a given lead compound in QSAR. The selection is performed such as to maximize the information carried by the descriptors representing the physicochemical features of the substituents. The maximum information is obtained by selection of substituents in a multivariate, variance-maximizing way. The obtained design matrix shows then quasi-orthogonal features that can be reliably used in regression models, where dependent variables, like biological activity, have to be modeled.

INTRODUCTION

In quantitative structure–activity relationship studies (QSAR), compounds are designed by varying a number of substituents around a skeleton of a "lead" compound in order to maximize a specific biological activity. The selection of substituents is traditionally based on the experience or the intuition of the researcher. More often trial-and-error attempts are used to localize promising classes of substituents able to induce biological potency in the test compounds. However, scarce attention is paid to the optimization of this procedure. Two main flaws are discernible in scrutinizing the majority of drug design studies. First, substituents are changed following the obsolete principle of "one-at-a-time", the monovariation method. That means that all substituents $R_1, R_2, R_3 \ldots R_n$ of a given lead compound L are kept the same except one, R_k, which is varied m times to give rise to m analogs. Too many similar compounds with small informational content are thus synthesized.

Second, when multivariation occurs, i.e., when two or more substituents R_i, R_k, R_j ..., are simultaneously varied around the constant frame of L, it often happens that the parameters describing the various R values show a high degree of covariance, limiting again the informational content of the analogs. As a consequence, the data matrix which is produced by the experiments has only little variance in some variables and contains highly redundant information.

In these cases the statistical analysis of the experimental data cannot give satisfactory information on the effect of substituent properties on the activity of a compound. This information is not contained in the data because of an improper experimental design. The proper design would require maximization of the variance of the substituent's data matrix. Promising attempts to select substituents in an information-maximized way have been based on the principal component analysis of some features of several organic substituents.[1]

We have attempted to develop a computer program which generates an information-maximized design if a lead compound is known, and the positions of the substituents are given working on the direct variables' (descriptors) space avoiding any principal component preprocessing.

It suggests for each experiment (test compound) the substituents which give the highest variance in the final data matrix. The selection of substituents in this approach is based on the known properties of each substituent such as inductive and mesomeric effect, size, lipophilicity, and polarizability.

The program is easy to use and requires almost no chemometrical knowledge for chemists who uses it. A preliminary presentation of ANALOGS was given elsewhere;[2] here the full philosophy and algorithm is presented.

THEORY

In an n-dimensional variables space, P, m objects, S, are described by m sets of coordinate n-tuples. A certain object $S_k = S(x_{k1}, x_{k2}, \ldots x_{kn})$ is uniquely located in the pattern space P and has a definite Euclidean distance to other objects S_j embedded in P. Each S_k is, therefore, easily represented by a vector $\mathbf{s}_k$. Any two objects, S_k and S_j, can be located in P such that their corresponding vectors $\mathbf{s}_k$,$\mathbf{s}_k$ show a varying degree of covariance. If $\mathbf{s}_j = c\mathbf{s}_k$, then $\mathbf{s}_j$,$\mathbf{s}_k = 1$, where c is any constant, and the vectors are collinear. If $\mathbf{s}_j$,$\mathbf{s}_k = 0$, they are orthogonal. Any value between 0 and 1 tells about the degree of covariance between the two vectors. Now, with m objects and n variables we have an $n{\cdot}m$ matrix **M** describing the spatial distribution of objects in P. **M** allows us to calculate the mutual distances between objects by the equation $d_{kj} = \sqrt{(d_k^2 - d_j^2)}$. This is one possible measure of similarity between objects. The covariance matrix **C** of the independent variables x_i is given by $C = \mathbf{M'M}$, where $\mathbf{M'}$ is the transpose of **M**. If the vectors **d** are normalized, then the correlation matrix **R** is obtained. Information theory shows that the total information I obtainable from a system described by a certain variable x that can have i values is given by

$$I(x) = -p(x)_i \ln p(x)_i$$

where $p(x)_i$ is the probability of existence of the ith value of x. Extension to more variables is trivial. It is evident that increasing the number of different states i for x augments the total information $I(x)$.

Further it can be shown that I is also dependent from C (or R) as follows

$$I = \text{const} \ln \det|C|$$

Suppose, as a simple example that C is given by the following two-by-two matrix

$$C = \begin{pmatrix} a & c \\ c & d \end{pmatrix}$$

The determinant of C is given by $\det|C| = ad - cc = q$, and $I = \text{const} \ln q$.

How Can I Be Maximized? If the off-diagonal elements c vanish, then the product of the diagonal elements a and d is left as the value for $\det|C|$, $ad - 0 = r$. As r is always larger

† University of L'Aquila.
‡ CHEMODATA.

0095-2338/93/1633-0266$04.00/0

than q, so $I(r)$ is also larger than $I(q)$. We see that diagonal matrices have maximum information. But diagonal matrices tell us that all the nondiagonal elements must be zero (or close to it as far as possible), meaning that the m variables of the original matrix **M** must be orthogonal! Orthogonality is the key to design matrices which offer maximum nonredundant and useful information for any later data processing.

In addition, orthogonal vectors show interesting properties concerning interactions among variables. Let x_1, x_2, and x_3 be three collinear vectors having, for simplicity, just two values, high (+1) and low (−1), describing eight objects. Due to their collinearity we have that $x_1 = cx_2$, $x_2 = c'x_3$ and, assuming $c = c' = 1$ for simplicity, we write a design matrix like

$$\begin{array}{ccc} x_1 & x_2 & x_3 \\ +1 & +1 & +1 \\ +1 & +1 & +1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \\ +1 & +1 & +1 \\ +1 & +1 & +1 \\ -1 & -1 & -1 \\ -1 & -1 & -1 \end{array}$$

The product of the eight vector elements of x_1,x_2 generates the eight vector elements of a new vector x_4, representing the interaction between x_1 and x_2.

We find $x_4 = (+1, +1, +1, +1, +1, +1, +1, +1)$. The sum of the eight elements of x_4 is nonzero, as x_2 is not independent from x_1. In fact, x_4 shows only one value, +1, and is thus a constant. The proof is given by calculating another vector x_5 by the product of the elements of x_1,$x_4 = (+1, -1, +1, -1, +1, -1, +1, -1)$, which is equal to x_1!

The effect of an interaction term of two collinear variables is not computable. The same holds for any pair of column vectors above. Let us use two orthogonal vectors x_1 and x_2 having the same two values (high,low) describing again eight objects. The matrix is obtained by taking all the possible multivariate permutations of the +,− signs

$$\begin{array}{ccc} x_1 & x_2 & x_3 \\ +1 & +1 & +1 \\ -1 & +1 & +1 \\ +1 & -1 & +1 \\ -1 & -1 & +1 \\ +1 & +1 & -1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \\ -1 & -1 & -1 \end{array}$$

It can be easily checked that any two column vectors are mutually orthogonal ($x_j{\cdot}x_k = 0$) and that the eight elements of any interaction vector x_4 are not a constant, for example, $x_4 = x_1{\cdot}x_3 = (+1, -1, +1, -1, -1, +1, -1, +1)$. The sum of all elements is 0, being x_1 orthogonal to x_3, and x_4 is a different vector from x_1,x_2 and x_3. This shows, in principle, that interaction effects become apparent when one deals with orthogonal arrays.[3] Orthogonal matrices have the very property that we want in order to run experiments with maximum informational content: their covariance matrix has nonzero diagonal elements, whereas all off-diagonal elements are zero. This leads to a maximum value determinant of C.

From all that has been explained so far, a few basic rules appear to satisfy our request for optimum design of sets of experiments:

1. Values for variables must stretch over a wide range (increases the magnitude of elements of the covariance matrix).
2. As many values (levels) as possible for each variable should be chosen (mapping density of pattern space increases, increase in I).
3. Multivariation of variable levels using orthogonal designs maximizes variance of the experimental matrix and gives access to interactions.
4. Multivariate matrix designs allow symmetrical, unbiased exploration of pattern space.

Where is Chemistry? The leap to chemistry is easily fulfilled by calling the above objects (the rows of the matrices) "analog molecules" and seeing the variables x as parameters (descriptors) which describe physical and chemical properties of the substituents. Of course, a chemical substituent cannot be "orthogonal" to another substituent, as nobody reasonably can call a methoxy group orthogonal to a methylamino group. What can be orthogonal is just the mathematical description of substituents, that is, the vectors we, the chemists, freely agree to use as representing some feature of a substituent like the methoxy group.

On the basis of the above considerations of mathematical nature, we tried to convert the methodology of abstract matrix design into an algorithm that would autodeductively select sets of chemical substituents in a way such that their corresponding descriptor matrix would have a maximum degree of intrinsic information. This is all we can do at the beginning of a drug design study, where a vector of responses y, the biological measurements for each tested analog, has to be modeled by the original, independent descriptors. The higher the intrinsic variance among the test compounds, the higher their degree of multivariation, the higher is the chance of understanding the real, independent effects of the selected descriptors upon the response variable. Especially in the beginning stage of a QSAR study, when one looks for crude, primary effects like "...will increasing inductive effect on L alter the bioactivity, and if yes in which direction?", it seems useful to start off with a minimum set of variance-maximized analogs.

Methods using D-optimal design[4] have been proposed[1] where the ensemble of substituents are described by a number of traditional QSAR descriptors and then projected onto a principal component (PC) space. In this PC space, the substituents cluster according to certain similarities within the describing features. D-optimal designs are then used to select that particular subset of substituents for which the determinant of C is maximized. This procedure, although very flexible, is quite time-consuming and requires PC preprocessing and D-optimal algorithms for delivering a solution. The method used in ANALOGS uses the direct, untransformed feature space spanned by the descriptors selected by the researcher that he feels is relevant for his QSAR study.

METHODS

(A) Molecular Structure. The skeleton of the lead compound can be drawn on a graphics terminal using a mouse. Direct input of a traditional connectivity matrix is equally possible if no graphics facilities are available. The positions of the variable substituents must be marked as R_1, R_2, etc. All other information required by the program are provided during the interactive session.

The program perceives the given molecular structure and detects possible interactions of substituent effects (interaction is understood here in a chemical, not in a mathematical sense). For example, two substituents connected to one delocalized π-electron system may have the same mesomeric effect on the whole system if they are in a certain symmetric position, like meta-substituents on benzene rings. Thus, to avoid redundancies (collinearity), one variable less is taken for the design. ANALOGS recognizes these situations and issues a warning message. The chemist may then decide whether to include this additional information into the design or not.

The user can tell the system that the properties of two substituents have the same effect on the activity, e.g., if it is known that electron-withdrawing substituents in both positions R_1 and R_2 make the compound more active, this information can be used to reduce the number of variables in the design.

(B) Substituents. ANALOGS uses an external file (RAWFILE) loaded with data on substituents and their properties. This file contains a list of common substituents, linked to a number of substituent descriptors like σ-para, σ-meta, π, and other calculated data, e.g., polarizabilities and charges. Several sets of substituents are available with different combinations of substituent properties, as not all data are available for all substituents. A typical file contains approximately 100 different substituents. The file organization is simple so that each user can easily modify or extend the substituent sets.

(C) Substituent Properties. Different substituent files contain the following sets of substituent properties:

(a) σ-para, σ-meta, π, ES

(b) σ-para, σ-meta, π

(c) Q-σ, Q-π, polarizability

Q-σ, Q-π, and polarizability are calculated with the PETRA software package.[5] Q-σ is the net charge produced on carbon atom 1 of a vinyl group connected to the substituent, and Q-π is the π-charge produced on carbon atom 2 of a vinyl group connected to the substituents. The substituent constants σ-para, σ-meta, π, and ES are taken from the tables by Hansch and Leo.[6]

Due to the simple file organization, users can add more substituent properties to the data set. During the program run, ANALOGS asks which of the properties on the file should be considered in the design. In this way the user can include additional experience he has in the problem at hand. For example, if it is known already that mesomeric effects have no influence on the activity of a given compound, they may be excluded from the design in order to reduce the number of variables in the data matrix.

It is important at this point to understand that the user creating his own RAWFILE should use variables that somehow do carry significant information, avoiding trivial covariances: using molecular volume and molecular refractivity as two hypothetical RAWFILE variables would end into insignificant designs as they are highly covariant. Only variables carrying independent information should be used to span the search space. This is a very easy task to do even if no previous inspection of a correlation matrix of the chosen variables is performed, just using chemical common sense.

(D) Experimental Design. As the first step in the design, ANALOGS inquires how may experiments (i.e., test compounds) the user can afford. This number is used as a guideline to work out the actual number of rows of the final design matrix.

If the number provided by the chemist is smaller than the number of variables m, the program gives a message containing the minimum number of experiments and asks for a new number. Next, the program suggests to do a factorial or fractional factorial design[7] with the nearest number of experiments (2 to the power of k). If the user agrees, the design is generated. Otherwise a design with half the number of compounds is suggested.

If this procedure yielded an acceptable number of experiments, the (fractional) factoral design is produced in two steps: generation of full factorial design in k variables and derivation of the remaining $m - k$ variable settings as products of the k variable settings. The result is a design matrix with 2^k rows and m columns. Each entry in this matrix is either $+1$ or -1, representing a "high" or a "low" value of the variable (substituent property). Which values of a substituent descriptor correspond approximately to "high", "low", and "medium" is automatically determined from the range of values in the substituent data file.

If, due to too many substituents and too many properties, a fractional factorial design is not feasible (e.g., the required testing would be too expensive), ANALOGS suggests a very simple design with $m + 1$ experiments which gives only trends of the variable effects. For the first compound, all substituent properties are set to "medium" values. For the following experiments in each compound one property is set to "high" and the others to "medium".

Nonlinear designs (composite design) are also possible. Here each descriptor is given more than two (+,–) values, up to five. They are particularly useful when the researcher suspects that the behavior of the dependent variable is not linearly linked to the descriptors, but shows a curvature.

(E) Selection of Substituents. For each experiment (analog) in the design, substituents must be selected from the external file to fit the design pattern (e.g., inductive effect high, mesomeric effect low, and high lipophilicity, etc.). The following procedure was chosen to find the appropriate substituents: The substituent properties on the file are scaled to a mean value of 0 and a variance of 1 (autoscaling). For each substituent required in the design an "optimal phantom substituent" is generated with property entries $+10$ and -10. The ranges of all phantom substituents delimit the shape and extension of our pattern space L, which therefore must contain all real substituents coded within the file. The distance of each real substituent in the file to this optimal substituent is calculated. After that the list of substituents is rearranged with respect to the distance to the optimal substituent. The substituent with the lowest Euclidean distance is selected. This is necessary because an ideal matrix design, for example, the two full factorial designs discussed in the previous section, spans a perfect cube in a 3D space. Obviously, empirical parameters like lipophilicity or charges cannot be located in L such that they build a perfect, symmetric polyhedron, because their actual values are given a priori for each chemical substructure and show fixed, discrete values (in contrast to pH, temperature, pressure, etc. which can be arbitrarily defined). Thus, the best approximation of a polyhedron is searched, which would match the ideal, symmetrical phantom polyhedron of "optimal substituents". The substituents which have been selected for each test compound are shown to the user who must agree or disagree. It may happen that a particular substituent or combination of substituents is judged by the chemist as too expensive in synthesis or necessitating a much too difficult chemistry. If one substituent is rejected by the user, the next substituent in the sorted list is suggested. This is repeated until a complete set of substituents has been accepted for each test compound. These compounds are written to a protocol file, which can be used to synthesize the test compounds.

TECHNICAL DETAILS AND STORAGE OF SUBSTITUENT PROPERTY LISTS

The selection of substituents with a defined list of properties (pattern), e.g., electron withdrawing and high lipophilicity, from a list involves a number of special tasks:

(a) Scaling of property numbers to make them comparable (e.g., autoscaling).

(b) Calculation of the similarity of substituents (i.e., Euclidean distance in property space).

(c) Sorting of the list of substituents according to similarity to design pattern (highest similarity first).
(d) Selection of most similar substituent. Selection of next similar substituent if one was rejected by the user.

These steps may be very time-consuming if a long list of substituents, typically 50–100, is used. Especially the calculation of a distance matrix and the sorting of the substituents on the file may cause long waiting times for the user.

Thus, a specialized file storage system was developed for this application, which allows very fast access to the presorted list of substituents.

Steps (a)–(c) in the above table have been moved to a file installation system (INSTAL). This program reads the list of substituents and their properties from an ASCII file (RAWFILE), performs operations (a)–(c), and stores the resulting list on a direct access file. This file contains a sorted list of substituents for each design pattern which may occur during one ANALOGS design process. ANALOGS can then pick the appropriate substituents at any time by one simple file access operation. As a result, there are no waiting times for a user during substituent selection. The time-consuming calculations are all done during the file installation process, which is performed only once by INSTAL, before the run of ANALOGS.

The access to the direct access file (BIGFILE) by ANALOGS is by index numbers. If ANALOGS has to select a substituent most similar to a design pattern calculated, it transforms this pattern into an index number. ANALOGS is written in FTN77 and has been implemented on SUN workstations and IBM PCs (without graphics).

In the Appendix, one sample output of a simple run is given. A molecule having two substituents R_1 and R_2 at a connecting double bond has been input. The molecule is unsymmetrical, so that mesomeric effects do act differently on the whole molecular body. During the interactive run, the user rejects specific substituents as unsuitable for synthesis and has them replaced by the next similar ones. The further run specifications by the user were selected parameters: π, σ-meta, σ-para; number of molecules the user wants to synthetise: maximum 10.

The program reduces them to 8, which correspond to a (6–3) fractional factorial design. 6 is generated by three descriptors and two positions. A full factorial design would require $2^6 = 64$ patterns. However, by the user-given constraint of 10 molecules, the next best design is one with 8 patterns.

Rejected substituents: -C≡C-H in position 1 at subsets 2 and 6, replaced by -CH_2Cl. Note: The output subsets are the results of the actual RAWFILE allocated substituents. Larger or smaller arrays of substituents, of course, may lead to different selection patterns.

CONCLUSIONS

In order to avoid redundancy and lack of valuable information in designing analogs of a certain lead compound, and aiming at reducing time and costs of synthetic work, one has to design a set of molecules such that the properties of the varying substituents show maximum variance. According to this principle, which guarantees for intrinsic maximum information in the data matrix, the method of multivariation should be conveniently used. Multivariation of the substituent descriptors generates a minimum number of test compounds to be created, which nevertheless show a maximum effect of their descriptors on a measured dependent variable, usually the biological activity (other properties of materials are investigable, too, in principle). When using any regression method for calculating the importance of physicochemical parameters of a response y, it is convenient, for a correct interpretation of the variables' loadings, that the independent variables are kept as orthogonal (independent) as possible. The ANALOGS program helps in generating sets of substituents which are mutually as distant as possible within an accepted, direct pattern space of m chosen descriptors. This maximized expansion is equivalent to a maximum variance of the design matrix. ANALOGS cannot be understood as an instrument for finding "the active" compound or the new lead. This cannot be excluded, of course, but its general purpose is to localize main effects from a minimum synthetic effort. It may help in concluding that positive mesomeric effects are important, while lipophilicity is not. Fine structural tuning is an unavoidable second step to obtain the really best substrate, but we think it can be performed more easily if clearcut, unbiased, and condensed quantitative information is available to the experimentor.

APPENDIX

Experiment 1:
substituent position 1: -NH-CO-NH_2
substituent position 2: -CH=N-phenyl

Experiment 2:
substituent position 1: -CH_2Cl
substituent position 2: -O-phenyl

Experiment 3:
substituent position 1: -NH-SO_2-CH_3
substituent position 2: -C#C-phenyl

Experiment 4:
substituent position 1: -CH=N-phenyl
substituent position 2: -NH-CO-NH_2

Experiment 5:
substituent position 1: -CH_2-O-phenyl
substituent position 2: -CH_2-O-phenyl

Experiment 6:
substituent position 1: -C#C-phenyl
substituent position 2: -CH_2Cl

Experiment 7:
substituent position 1: -O-phenyl
substituent position 2: -NH-SO_2-CH_3

Experiment 8:
substituent position 1: -CF_2-CF_3
substituent position 2: -CF_2-CF_3

REFERENCES AND NOTES

(1) (a) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. *Quant. Struct.–Act. Relat.* **1989**, *9*, 32. (b) Baroni, M.; Clementi, S.; Cruciani, G.; Kettaneh-Wold, N.; Wold, S. submitted. (c) Clementi, S.; Cruciani, G.; Baroni, M.; Skagenberg, A. *Chim. Ind.* **1990**, *6*, 536.

(2) Marsili, M. *Tetrahedron Comput. Methodol.* **1988**, *1*, 1.

(3) The problem of confounding interactions with primary variables cannot be treated here. Refer to ref 7 for more information about this subject.

(4) (a) Mitchell, T. J. *Technometrics* **1974**, *16*, 203. (b) Mitchell, T. J.; Bayne, C. K. *Technometrics* **1978**, *20*, 369.

(5) PETRA Software Package for the calculation of physicochemical properties of molecules. CHEMODATA Computer-Chemie GmbH & Co. KG, Munich, FRG.

(6) Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

(7) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters*; John Wiley & Sons: New York, Chichester, Brisbane, and Toronto, 1978; pp 374–409.