Chemometric Methodologies in a Quantitative Structure−Activity Relationship Study: The Antibacterial Activity of 6-Aminoquinolones

Violetta Cecchetti,[†] Enrica Filipponi,[†] Arnaldo Fravolini,*[,†] Oriana Tabarrini,[†] Daniela Bonelli,[‡]
Monica Clementi,[‡] Gabriele Cruciani,[‡] and Sergio Clementi*[,‡]

Istituto di Chimica e Tecnologia del Farmaco, Università di Perugia, Via del Liceo 1, 06123 Perugia, Italy, and Laboratorio di Chemiometria, Dipartimento di Chimica, Università di Perugia, Via Elce di Sotto 8, 06123 Perugia, Italy

Received May 30, 1996[⊗]

The paper illustrates the chemometric strategies appropriate for extracting information from a large amount of biological data regarding the antibiotic activity of 6-aminoquinolones. The unique framework based on principal component analysis, projection onto latent structures, and response surface methodologies permits the structure−activity correlations to be shown and to suggest new compounds for further testing. The low activity of the suggested molecules points out the limitations of quantitative structure−activity relationship models when the training set is not properly designed in order to balance all the structural variations taken into account.

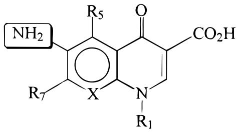
Introduction

Quinolones are a major class of antibacterial agents. They are still being developed to increase their activity and broaden their antibacterial spectrum. In two recent studies directed toward Gram-negative[1] and Gram-positive[2] bacterial strains, we showed how computer chemistry methodologies, namely the chemometric tools used in quantitative structure−activity relatioships (QSAR), allowed a rationalization of the structural features affecting the activity of the existing molecules with a subsequent suggestion of a new class of potentially active compounds.

The most relevant information obtained from these studies was that an amino group at C-6, instead of the the usual fluorine atom, still gives interesting and active molecules. A subsequent study undertaken on a whole class of 6-aminoquinolones has shown that these compounds are characterized by a good activity level and a rather broad antibacterial spectrum.[3,4] Because of the good MIC values of these compounds, we carried out a QSAR study in order to try to optimize the activity level on a broad antibacterial spectrum as well as to show a real example of how chemometric methodologies based on principal component analysis (PCA), projection onto latent structures (PLS), and response surface (RS) analysis allow information to be extracted from a large set of existing biological data, from which the pharmacological properties can be rationalized. The results can then suggest the subsequent direction which the research should take.

Materials

Molecules. The study was performed on a set of 39 derivatives synthesized and tested by our group. The molecules considered have a common basic *3-carboxy-6-amino-4-quinolone* or *3-carboxy-6-amino-4-naphtyridone* structure with various arrangements of the functional groups at positions 1, 5, 7, and 8 (Figure 1). The fluoroquinolones, ciprofloxacin and rufloxacin,[5a,b] were added for comparative purposes.



X = CH, N, CF, CCH_3
R_1= *c*-Pr, *t*-Bu, 4-FC_6H_4
R_5= H, NH_2
R_7 = see Table 5

Figure 1. General structure of the compounds under investigation.

Biological Activities. The series of quinolone and naphthyridone acids used in this study were tested *in vitro* against eight Gram-negative and five Gram-positive bacterial strains (Table 1). The activity level is estimated in terms of minimum inhibitory concentration (MIC) in μg/mL which is measured by the conventional agar dilution procedure.[5a] According to this method, the lowest antibiotic concentration able to inhibit the growth of a well-known quantity of micro-organisms is measured starting from an initial inhibitor concentration that is halved at each step. However, it is appropriate that the actual MIC values are not used as such, but transformed into a linear numerical scale (between 0 and 13: Table 2) that corresponds to taking the $\log_2$ MIC minus a constant term. This transformation allows the same importance to be attribute to each concentration halving. If the MIC values are used as such, the bad values (highest values of MIC) will take the lead in the resulting models. A $\log_{10}$ MIC transformation would also equally reduce the leverage of the bad values, but has no experimental basis.

Methods

Principal Component Analysis (PCA). Principal components analysis is a multivariate statistical analysis method[6] which permits a table of numbers to be transformed into a few informative diagrams, thus permitting a simple and straightforward interpretation of the problem under investigation. If we have a data matrix with N objects (molecules) described by P variables (the MICs against the different bacterial strains), each molecule can be considered as a point in a P-dimensional space. The main objective of a PC analysis is to find the lowest dimensionality model which can adequately describe the structure of the multivariate data. The model obtained can be regarded as a projection of the problem onto a space of reduced dimensionality. The coordinates of this space, called principal components are new directions in the original space which can be described as linear combinations of the original variables.

[*] Author to whom correspondence should be addressed.
[†] Istituto di Chimica e Tecnologia del Farmaco.
[‡] Laboratorio di Chemiometria, Dipartimento di Chimica.
[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1997.

S0022-2623(96)00385-8 CCC: $14.00

Table 1. Selected Bacterial Strains

Gram-negative organisms	Gram-positive organisms
(1) *Escherichia coli* ATCC 8739	(9) *Staphylococcus aureus* MPR 5
(2) *Escherichia coli* ISF 432	(10) *Staphylococcus aureus* ATCC 6538
(3) *Enterobacter cloacae* OMNFI 174	(11) *Staphylococcus epidermidis* HCF Berset C
(4) *Acinetobacter calcoloaceticus* OSMPV 113	(12) *Staphylococcus epidermidis* CPHL A2
(5) *Providencia stuardii* CNUR 5	(13) *Streptococcus faecalis* LEP Br
(6) *Klebsiella pneumoniae* ATCC 10031	
(7) *Shigella enteritidis*	
(8) *Pseudomonas aeruginosa* ATCC 9027	

Table 2. MIC Values and Their Transformations into ($\log_2$ MIC + C) and Desirability Values

MIC (μg/mL)	$\log_2$ MIC	desirability
128	0	0
64	1	0
32	2	0
16	3	0
8	4	0.1
4	5	0.2
2	6	0.3
1	7	0.4
0.5	8	0.5
0.25	9	0.6
0.12	10	0.7
0.06	11	0.8
0.03	12	0.9
0.01	13	1

The mathematical expression of a PCA model describes each element x_{ik} of the data matrix by the following equation (eq 1)

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia} p_{ak} + e_{ik} \qquad (1)$$

where *i* indicates the object (molecule), *k* indicates the variable (assay), and *A* is the total number of principal components "*a*" required by the model. The loadings, p_{ak}, represent the coefficients of the variables *k* in the linear combinations of the original variables that define each principal component "*a*", while the scores t_{ia} represent the position of the projection of the objects onto these new variables, and e_{ik} are the residuals.

Linear PLS Modeling. The purpose of a QSAR study is to find a statistical model capable of describing the biological activities (*y*'s) of a series of molecules in terms of a certain number of *x* variables which describe their chemical structures. The objective of the analysis is to provide a causal relationship between the **Y** vector and the **X** matrix so that the biological behavior of the already available series of molecules can be explained and the activity of compounds not yet synthesized and tested can be predicted. PLS (projection onto latent structures)[7a,b] has been shown to be the most appropriate regression method to derive QSAR models, whereas ordinary regression methods like multiple linear regression (MLR) might be misleading because of problems such as multicollinearity and fixed dimensionality.[8a,b]

A PLS model describes the **X** matrix by a principal component-like model (eq 1) and the **Y** vector as a predictive relationship with the principal components, here called latent variables (eq 2, where b_a is a proportionality coefficient for each dimension *a*), under the constraint of maximizing the correlation between *y* and *t*.

$$y_i = \bar{y} + \sum_{a=1}^{A} b_a t_{ia} + f_i \qquad (2)$$

The results of a PLS model allow the relative importance of the structural features affecting the biological activity to be ranked on the basis of the relative importance of the *p* values, the loadings, which describe how much each original variable, here structural feature, participates in the definition of the latent variables. All the data analyses were carried out using the SIMCA method and package.[9,10]

Nonlinear PLS Modeling. Supposing that the relationship between the *y* values and the *x* variables is not linear, a QSAR problem can be also formulated so that the biological activity can be expressed as a response surface, which is mathematically described by a second-degree polynomial expression that, in the simplest case of two variables, assumes the form of eq 3.

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{11} x_1^{\ 2} + b_{22} x_2^{\ 2} + b_{12} x_1 x_2 \qquad (3)$$

The coefficients of the polynomial expression can be computed by multiple regression analysis if the experiments have been chosen according to an experimental design strategy such that the descriptor variables are independent and define a rotatable design. However, when variables are not independent, as happens with structural descriptors because of the discrete nature of organic molecules and substituents, the CARSO (computer-aided response surface optimization)[11] procedure allows the coefficients of the response surface to be computed by using the PLS algorithm.

The procedure is based on a linear PLS model built on an expanded matrix of descriptors which, besides the linear terms, also contains the squared terms and the bifactorial interaction terms of the variables. The loadings of the PLS model are transformed thereafter into the coefficients of the polynomial expression by simply setting the mathematical expressions of the linear and the quadratic model equal. Finally, the response surface is studied using the canonical and Lagrange analyses.

The CARSO procedure is particularly useful for formulating QSAR studies as optimization problems which allows us to find which substituent for each position considered could lead to a molecule with a sufficiently high activity. It should be mentioned that the purpose of the CARSO procedure is not to find a better description of the available data, but rather is to detect the interchangeable substituents for each substitution site.

Results and Discussion

PCA on the Biological Activity Matrix. PCA is the most appropriate tool for extracting the systematic information contained in a data matrix. On performing a PC analysis on the data of Table 3, where molecular structures are coded according to Tables 4 and 5, we obtained a three-component model which explains 92% of the total variance (see Table 6). This means that we have three independent effects that contribute to the total variance, while the remaining 8% is due to the "noise" of the data. The number of significant components was assessed by the criteria embedded in SIMCA, i.e. in terms of model predictivity (Q^2 being 0.88 for A = 3 with no further increase).

The first component explains 78% of the total variance, and all the variables have almost the same contribution (all the loadings of Table 6 are positive and quite similar). This component is related to the antibiotic potency against the bacterial strains. Since all the considered molecules are active against all the

Table 3. *In Vitro* Antibacterial Activity ($\log_2$ MIC)[a].

objects		variables												
		1 *E. co.*	2 *E. co.*	3 *E. cl.*	4 *A. ca.*	5 *P. st.*	6 *K. pn.*	7 *S. en.*	8 *P. ae.*	9 *S. au.*	10 *S. au.*	11 *S. ep.*	12 *S. ep.*	13 *S. fe.*
1	00	6	6	3	0	1	6	3	0	0	0	1	0	0
2	1a	9	12	9	5	4	12	9	5	6	6	5	6	2
3	1b	9	12	9	3	2	9	8	6	1	1	3	3	0
4	1c	12	12	9	3	3	12	9	6	3	3	5	5	2
5	1d	12	12	8	4	5	12	9	6	4	4	5	6	2
6	1e	5	6	6	1	1	6	6	2	1	1	1	1	1
7	1f	8	8	6	2	0	8	6	2	0	0	2	1	0
8	1g	12	12	8	7	7	12	8	6	8	8	6	7	4
9	1h	12	12	6	6	4	12	6	4	5	5	5	6	4
10	1i	9	9	4	0	0	9	3	0	0	0	0	0	0
11	1j	5	5	0	0	0	5	0	0	0	0	0	0	0
12	1k	7	8	3	0	0	7	3	3	0	0	0	0	0
13	1l	8	7	3	0	0	8	2	0	0	1	1	1	0
14	1m	6	5	1	0	0	6	1	0	0	0	0	0	0
15	1n	7	7	0	0	0	6	0	0	0	0	0	0	0
16	2a	11	11	7	5	5	11	7	5	6	6	5	6	2
17	2g	11	12	6	8	5	12	6	5	9	9	8	8	4
18	2i	10	10	3	3	1	10	2	1	4	5	4	4	0
19	2o	10	10	4	4	5	10	4	2	6	6	5	5	1
20	2p	10	10	4	6	5	10	4	2	8	9	7	8	4
21	2q	8	8	1	7	0	9	1	0	7	7	6	7	3
22	3a	10	10	5	3	5	10	5	4	5	4	4	4	1
23	3g	11	12	5	5	5	12	5	0	7	8	5	5	0
24	4a	10	12	10	5	6	12	10	6	5	5	5	5	0
25	5a	7	11	9	4	5	11	8	6	5	5	4	5	1
26	5g	10	11	7	7	8	12	5	5	9	8	7	8	5
27	6a	9	13	9	8	8	13	8	4	6	6	6	6	4
28	7a	3	3	0	0	0	2	0	0	0	0	0	0	0
29	8a	12	12	10	6	6	12	9	5	6	6	5	5	5
30	C[b]	12	13	13	11	11	12	12	11	10	10	10	11	8
31	9a	12	12	10	–	6	12	11	6	8	8	7	7	5
32	9g	9	12	9	10[c]	6	12.5	10	7	11	11	10	10	8
33	9c	10	12	10	6[c]	6	12	11	7	7	7	7	7	5
34	9i	2	7	1	3[c]	1	8	1	0	5	5	3	3	1
35	9o	13	13	9	–	11	11	9	8	13	13	12	13	9
36	9r	12	12	10	–	8	12	10	7	10	10	10	9	7
37	9h	7	12	8	10[c]	7	12	9	6	10	11	11	11	8
38	9j	10[c]	10[c]	5[c]	5[c]	2[c]	9[c]	5[c]	3[c]	3[c]	3[c]	3[c]	3[c]	1[c]
39	9d	11	12	8	6[c]	6	12	9	6	7	7	6	7	3
40	9b	12	12	11	6[c]	6	12	12	7	5	6	4	5	3
41	C[b]	13	13	13	10	11	13	13	11	11	11	11	11	8
42	R[b]	9	9	8	7	6	9	10	5	8	8	7	8	5

[a] See refs 3 and 4 unless otherwise stated. [b] C (ciprofloxacin) and R (rufloxacin) are the reference compounds. Molecules 30 and 41 are different measurements of biological activity for the same compound. [c] Unpublished data.

Table 4. Subclasses of the Molecules Considered

code	substituents			
	R_1	R_5	R_7	X
class 0	*c*-Pr	H	Cl	CH
class 1	*c*-Pr	H	heterocycle	CH
class 2	*t*-Bu	H	heterocycle	CH
class 3	4-FC_6H_4	H	heterocycle	CH
class 4	*c*-Pr	H	heterocycle	N
class 5	*c*-Pr	H	heterocycle	CF
class 6	*c*-Pr	NH_2	heterocycle	N
class 7	*c*-Pr	NH_2	heterocycle	CH
class 8	*c*-Pr	NH_2	heterocycle	CF
class 9	*c*-Pr	H	heterocycle	C-CH_3

bacterial strains, the first component ranks molecules according to potency and is not informative about the real differences between molecules and/or micro-organisms.

The second component explains a further 9% of the total variance. In this case different groups of variables contribute in different ways in determining the component, as seen from the second column of Table 6 and from the loading plot (Figure 2) of the first *versus* the second component. The latter shows that there is a marked separation between Gram-positive (upper part of the plot) and Gram-negative (lower part) bacterial strains.

It should be noted that variable 5 (*P. stuardii*) is in the middle of the plot and therefore provides little information. Moreover, variable 4 (*A. calcoloaceticus*), which represents the activity of a Gram-negative bacterial strain, has the same behavior as the other Gram-positive bacterial strains considered. We suggest that this strain may not be well classified.

The corresponding score plot of the first *versus* the second component (Figure 3) shows that the molecules are ordered in a very informative way: (a) the antibiotic activity against the considered micro-organisms increases from left to right; (b) molecules that are preferably active against the Gram-positive bacterial strains are found in the upper part of the plot, while those preferably active against the Gram-negative ones are in the lower part.

For an easier graphic interpretation of the chemometric results, molecules have been coded by a number and a letter; the number indicates the subclass of molecules according to Table 4, while the letter indicates the side chain sitting at the C-7 position as reported in Table 5. Using these codes we can easily rationalize

Table 5. Heterocyclic Side Chains Employed as R_7 Substituent and Relative Codes

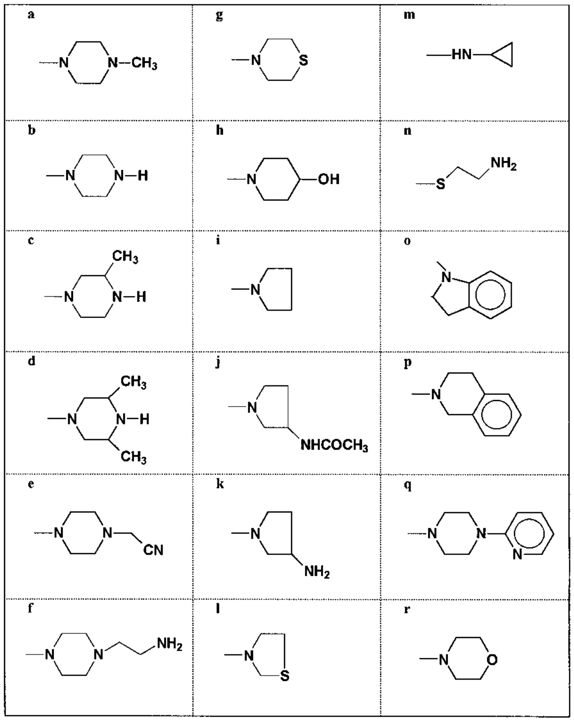


Table 6. Loadings of the Principal Component Analysis[a]

	PCA Loadings		
variables	PC1 (p1)	PC2 (p2)	PC3 (p3)
var 1 (*E. coli*)	+0.23	−0.32	−0.50
var 2 (*E. coli*)	+0.28	−0.27	−0.32
var 3 (*E. cloacae*)	+0.26	−0.38	+0.31
var 4 (*A. calcoloaceticus*)	+0.30	+0.23	+0.00
var 5 (*P. stuardii*)	+0.29	+0.02	+0.11
var 6 (*K. pneumoniae*)	+0.27	−0.21	−0.40
var 7 (*S. enteritidis*)	+0.26	−0.35	+0.35
var 8 (*P. aeruginosa*)	+0.27	−0.24	+0.40
var 9 (*S. aureus*)	+0.28	+0.31	−0.10
var 10 (*S. aureus*)	+0.28	+0.32	−0.13
var 11 (*S. epidermidis*)	+0.29	+0.24	+0.00
var 12 (*S. epidermidis*)	+0.29	+0.24	−0.03
var 13 (*S. faecalis*)	+0.27	+0.28	+0.27
explained variance	78%	9%	5%
cumulative variance	78%	87%	92%

[a] The + and - signs indicate the way each variable participates in each principal component.

the effects of different structural variations on the antibacterial activity. Compounds with a methyl group in the C-8 position (class 9) generally show a higher activity (Figure 4); in contrast, molecules with a hydrogen in the C-8 position (class 1) have a much lower antibacterial potency.

Similarly, when heterocycles like “o” (2-isoindolinyl), “p” (1,2,3,4-tethrahydro-1-isoquinolinyl), “q” (4-(2-piperidyl)-1-piperazinyl), or “g” (1-thiomorpholinyl) are in the C-7 position, the corresponding molecules have a higher activity level against Gram-positive bacterial strains (Figure 5). With substituents like “a” (1-methyl-1-piperazinyl), “b” (1-piperazinyl), “e” (3-metyl-1-piperazinyl), or “d” (3,5-dimethyl-1-piperazinyl), the compounds are mainly active against the Gram-negative ones. This interesting result can be attributed to different interactions of the substituents with Gram-positive and Gram-negative bacterial cell walls.

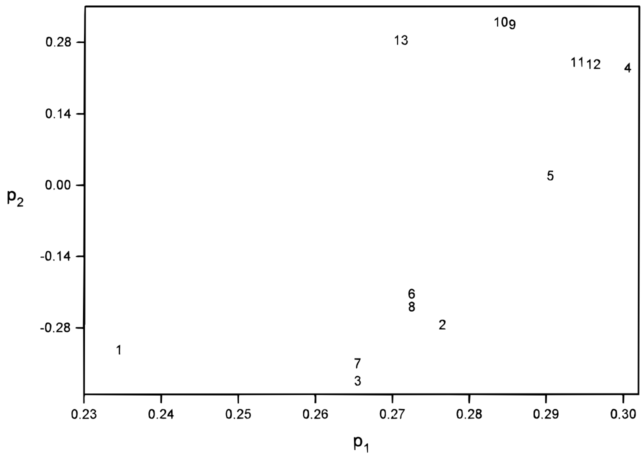


Figure 2. PCA loading plot of the first *versus* the second component. The numbers correspond to the bacterial strains.

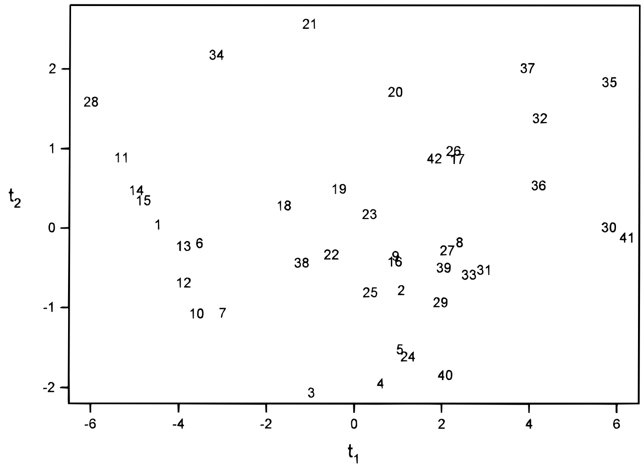


Figure 3. PCA score plot of the first *versus* the second component. The numbers correspond to the molecules.

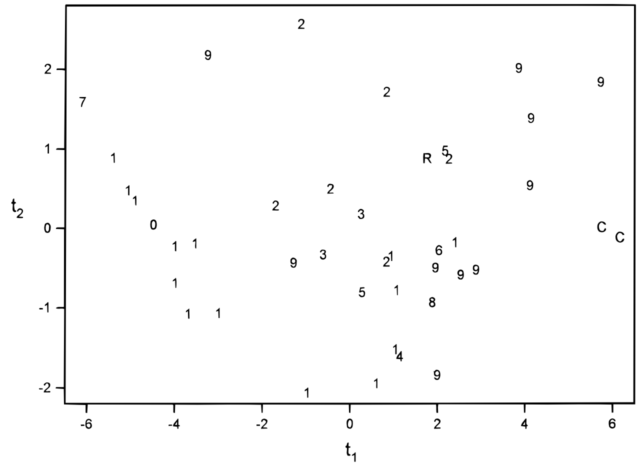


Figure 4. PCA score plot of the first *versus* the second component with the codes that indicate the subclass of molecules according to Table 4.

Finally, the third component explains a further 5% of the total variance. In spite of the low fraction of explained variance, the loading plot of the second *versus* the third component is quite informative (Figure 6). In fact, it allows three different groups of micro-organisms to be recognized that behave similarly when treated with the same antibiotics. The first group contains variables 1, 2, and 6 (two *E. coli* and *K. pneumoniae* bacterial strains): these Gram-negative bacterial strains

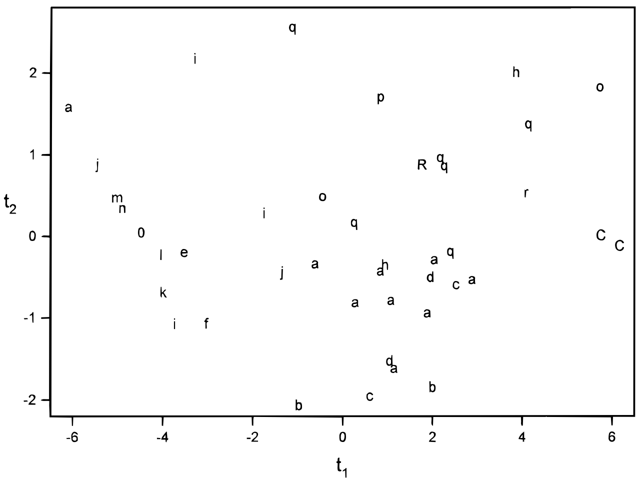


Figure 5. PCA score plot of the first *versus* the second component with the codes that indicate the side chain at position 7 according to Table 5.

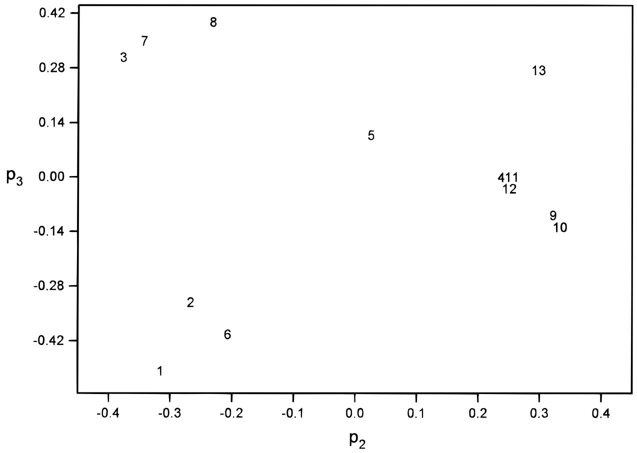


Figure 6. PCA loading plot of the second *versus* the third component. The numbers correspond to the bacterial strains.

are quite easy to inhibit. The second group is formed by the variables 3 (*E. cloacae*), 7 (*S. enteriditis*), and 8 (*P. aeruginosa*). These variables correspond to high MIC values. These three Gram-negative bacterial strains are therefore not easy to inhibit. The third group contains all the Gram-positive bacterial strains (variables from 9 to 13) and *A. calcoloaceticus* (variable 4). As already stated, the latter shows the same behavior as the group of Gram-positive bacterial strains, in spite of being known as Gram-negative. The plot also confirms that variable 5 (*P. stuardii*) does not fit the systematic information provided by other micro-organisms. It is as if it identifies a new group of bacterial strains.

Transformation of Responses into Desirability Functions. The objective of the second part of this work was to identify the structural features of quinolone derivatives that affect the antibacterial activity against all three groups of bacterial strains found by the previous PCA. On choosing just one bacterial strain as representative for each group, we can sum up the information using only three variables. On the basis of their diffusion and reliability, we have chosen *E. coli* (variable 1), *P. aeruginosa* (variable 8), and *S. aureus* (variable 10). This reduction of the variables does not significantly decrease the overall information because of the redundancy contained within each group of variables (Figure 6). Consequently, we need to consider three responses simultaneously in order to find the best compromise between them. In fact, our final aim is to rationalize the structural features that would increase the antibacterial activity against all types of bacterial strains.

With several responses it is possible to use the PLS2 algorithm, which would find the correlations between them, but is not aimed at finding the best compromise. To do this, an external evaluation of the goodness of these responses, referred to as expectations, is introduced by means of the application of desirability functions.[12] In fact, if the responses change in different or independent ways (e.g. potency and toxicity of a drug), a simple PLS analysis does not provide information about their best compromise.

A desirability function is a transformation function which permits a response to be modified in order to take into account its goodness. This function is defined as a dimensionless scale between zero and one. Zero is assigned to a response value which is not considered to be good enough for the expected property of the molecule, while one is attributed to a response value above which it is of little use to increase further the property. Assigning the values zero and one is therefore a subjective choice that should be based on the level of understanding the problem. The intermediate values between zero and one are obtained by means of the transformation function which can assume different shapes (linear, exponential, ...).

The desirability values for our antibacterial responses are reported in the last column of Table 2. In other words, when the MIC value is greater than 16 μg/mL, we assume that the compound is not interesting enough and we set the response equal to zero. On the contrary, when the MIC is smaller than 0.01 μg/mL, the compound has such a good potency that to increase it further would be irrelevant and we can set the response equal to one. All intermediate values correspond to a linear transformation from the minimum to the maximum value. The transformation of real biological data into desirability functions is particularly appropriate because it is possible to define a total desirability function as the geometric mean of the individual ones which represents in a suitable way the compromise between the different individual responses (eq 4).

$$D_{\mathrm{tot}} = (d_1 d_2 ... d_{\mathrm{n}})^{1/n} \qquad (4)$$

In fact, the total desirability also changes from zero to one; it is high only when all the individual functions are high and it is low or even goes to zero when just one of the functions is low or zero. By using the total desirability function, all the dependent variables can be summarized in one response only.

Structural Description of Quinolones. To describe the structure of the quinolone derivatives, a conventional QSAR procedure was used following the tradition developed by Hansch.[13] In fact, only the substituents sitting at different positions are described; the common quinolone basic structure is not taken into account.

The choice of the traditional descriptors to be used is always a subjective one. Although PLS is appropriate for handling a large number of descriptors, it is useless to do so when the number of varying substituents is low

Table 7. Data Matrix for Linear PLS Analysis[a]

objects		variables										
		1	2	3	4	5	6	7	8	9	10	11
		MR[b]	σ_p[c]	MW[d]	N al[d]	N ar[d]	Ht ≠ N[d]	ring at[d]	NH_2 OH[d]	t_1[e]	t_2[e]	D_{tot}[f]
1	**1a**	**13.53**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**0**	**0**	**0.33**
2	1b	13.53	0	85	2	0	0	6	0	0	0	0
3	1c	13.53	0	99	2	0	0	6	0	0	0	0
4	**1d**	**13.53**	**0**	**113**	**2**	**0**	**0**	**6**	**0**	**0**	**0**	**0.3**
5	1e	13.53	0	124	3	0	0	6	0	0	0	0
6	1f	13.53	0	128	3	0	0	6	1	0	0	0
7	**1g**	**13.53**	**0**	**102**	**1**	**0**	**1**	**6**	**0**	**0**	**0**	**0.51**
8	**1h**	**13.53**	**0**	**100**	**1**	**0**	**0**	**6**	**1**	**0**	**0**	**0.26**
9	1i	13.53	0	70	1	0	0	5	0	0	0	0
10	1k	13.53	0	85	2	0	0	5	0	0	0	0
11	1l	13.53	0	88	1	0	1	5	0	0	0	0
12	**2a**	**19.62**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**0**	**0**	**0.36**
13	**2g**	**19.62**	**0**	**102**	**1**	**0**	**1**	**6**	**0**	**0**	**0**	**0.46**
14	2i	19.62	0	70	1	0	0	5	0	0	0	0
15	2o	19.62	0	118	1	0	0	5	0	0	0	0
16	2p	19.62	0	132	1	0	0	6	0	0	0	0
17	2q	19.62	0	162	2	1	0	6	0	0	0	0
18	3a	25.36	0	99	2	0	0	6	0	0	0	0.19
19	3g	25.36	0	102	1	0	1	6	0	0	0	0
20	**4a**	**13.53**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**0**	**−3.14**	**0.35**
21	**5a**	**13.53**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**0.39**	**−1.01**	**0.29**
22	**5g**	**13.53**	**0**	**102**	**1**	**0**	**1**	**6**	**0**	**0.39**	**−1.01**	**0.41**
23	6a	13.53	−0.66	99	2	0	0	6	0	0	−3.14	0.26
24	7a	13.53	−0.66	99	2	0	0	6	0	0	0	0
25	8a	13.53	−0.66	99	2	0	0	6	0	0.39	−1.01	0.38
26	**9a**	**13.53**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**1.22**	**−0.21**	**0.51**
27	**9g**	**13.53**	**0**	**102**	**1**	**0**	**1**	**6**	**0**	**1.22**	**−0.21**	**0.58**
28	**9c**	**13.53**	**0**	**99**	**2**	**0**	**0**	**6**	**0**	**1.22**	**−0.21**	**0.48**
29	9i	13.53	0	70	1	0	0	5	0	1.22	−0.21	0
30	**9o**	**13.53**	**0**	**118**	**1**	**0**	**0**	**5**	**0**	**1.22**	**−0.21**	**0.79**
31	**9r**	**13.53**	**0**	**86**	**1**	**0**	**1**	**6**	**0**	**1.22**	**−0.21**	**0.63**
32	**9h**	**13.53**	**0**	**100**	**1**	**0**	**0**	**6**	**1**	**1.22**	**−0.21**	**0.46**
33	**9d**	**13.53**	**0**	**113**	**2**	**0**	**0**	**6**	**0**	**1.22**	**−0.21**	**0.46**
34	**9b**	**13.53**	**0**	**85**	**2**	**0**	**0**	**6**	**0**	**1.22**	**−0.21**	**0.48**

[a] The bold objects are the active molecules used for the final PLS analysis ($Y > 0$). The topological descriptors are as follows: MW, molecular weight; N al, number of aliphatic N atoms; N ar, number of aromatic N atoms; Ht ≠ N, number of heteroatoms different from N; ring at, number of terms of the ring directly linked to the skeleton; NH_2 OH, number of NH_2 or OH groups. [b] R_1. [c] R_5. [d] R_7 (topological descriptors). [e] R_8. [f] **Y**.

at a certain site. Accordingly, we decided to consider one single descriptor for positions 1 and 5, and focused our attention to positions 7 and 8, where the structural variation is larger.

On the basis of results obtained in previous work[1] the R_1 substituents were described by means their molar refractivity (MR) values while the R_5 substituents were described by their σ_p constants which reflect their inductive and resonance electronic effects. The R_7 substitution site was described by a series of six topological descriptors already used in previous works.[1,2] They report the main characteristics of the considered heterocyclic chain substituents: (1) molecular weight, (2) number of aliphatic *N* atoms, (3) number of aromatic *N* atoms, (4) number of heteroatoms different from *N*, (5) number of terms of the ring directly linked to the skeleton, and (6) number of NH_2 or OH groups. For the R_8 substituents, the two first principal properties of organic substituents[14] were used, namely t_1 for the steric effect and t_2 for the electronic effect, which serve as statistical descriptors.

The QSAR table is reported in Table 7, where in order to obtain a more balanced series of compounds, the following were excluded: the reference compounds, the only molecule belonging to class zero and those molecules bearing the substituents “j”, “m”, and “n” at C-7, because they induce low activity and because these substituents could not be parameterized by the six indices chosen in a congruent way. The resulting data matrix therefore contains 34 molecules (the objects), one dependent variable (the total desirability), and 10 descriptor variables, where the substitution sites R_1 and R_5 are described by one parameter, R_7 by six parameters, and R_8 by two parameters. In order to give each substitution site the same initial importance for the PLS analysis, the data matrix was modified by a block scaling, i.e. multiplying each variable by a correction term $1/\mathrm{SD}\sqrt{n}$ where n is the total number of descriptors that characterize the considered substitution site, while SD is the standard deviation of the variable.

Linear PLS Analysis. The linear PLS analysis of this matrix gives a two-principal-component model which explains 55% of the total variance of y (D_{tot}). In detail the first component explains 42% of the variance and is mainly a combination of MR for the R_1 substituent and t_1 for the R_8 substituent. The second component explains a further 13% of the total variance and depends almost entirely upon the MR parameter for the substituent in R_1.

The score plot of Figure 7 shows, however, that the data set is not homogeneous; in fact, inactive compounds ($D_{tot} = 0$) cannot be modeled together with active ones. This situation is well-known in QSAR as the “asymmetric case”.[15] Since loss of activity can be due to the lack of any of the key structural features, there is no reason to expect that inactive compounds can be modeled at all, while for the active molecules, the activity

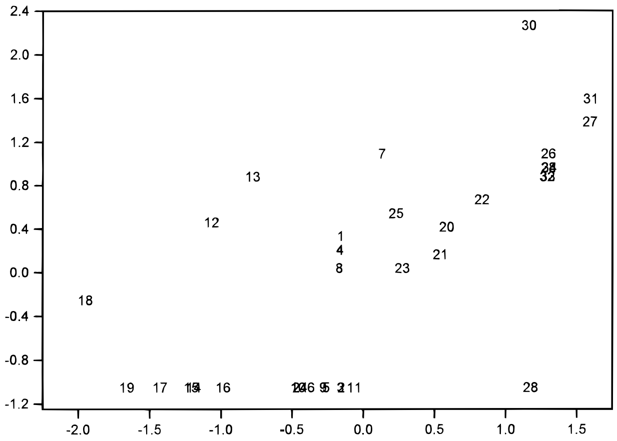


Figure 7. PLS score plot. All the objects on the bottom have $D_{tot} = 0$. The numbers correspond to the molecules.

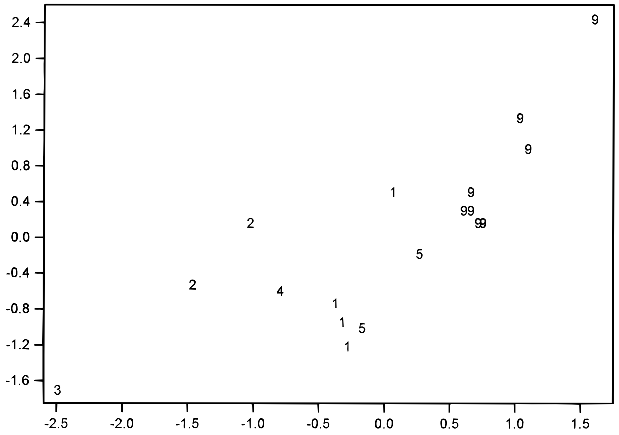


Figure 8. PLS score plot with the codes of Table 4.

variation parallels well-defined structural modifications. Although the exclusion of inactive compounds is a problem formulation requirement, it can be easily checked that they do not belong to the confidence space around the PLS model of the active ones in the descriptor space.

Therefore, the PLS analysis was repeated after excluding the 14 inactive molecules. We also excluded the two active molecules which bear an NH_2 group at C-5, because these would produce a skewed distribution of that descriptor, and therefore its apparent importance would be inflated. Consequently, the calculations were performed on a data matrix consisting of 18 objects, 10 descriptor variables, and one dependent variable. The new analysis also gives a two-component model, quite similar to the previous one, where the first latent variable explains 60% of the total variance and is a combination of MR for the R_1 substituent and t_1 for the R_8 substituent. The second latent variable explains a further 22% of the total variation of y and mainly contains the parameter MR for the R_1 substituent. The interpretation is therefore the same, even though almost half of the molecules were excluded and the model is much better from a numerical point of view.

It may be appropriate to observe that the variance explained by the PLS model cannot be compared with that explained by the previous PCA model, since the former refers to the D_{tot} vector and the latter to the raw data matrix of biological activities.

The PLS score plot is shown in Figure 8 with the codes for the nuclear substituents and in Figure 9 with the codes for the chain in the C-7 position. It is clear that molecules with a methyl group at C-8 (class 9) have by far the best activity and the broadest antibacterial spectrum, while the situation is not as clear for the R_7 and R_1 substituents.

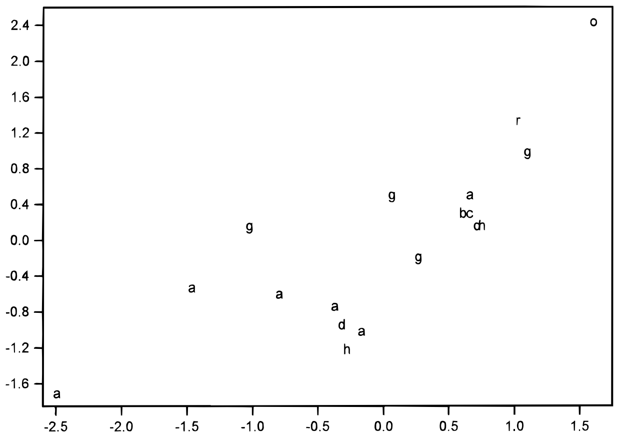


Figure 9. PLS score plot with the codes of Table 5.

Response Surface Study. Besides the information gained from linear PLS modeling, the QSAR problem can also be formulated as an optimization problem by means of a response surface study. In fact, response surfaces obtained by the CARSO procedure enable a data set to be rationalized in terms of nonlinear QSARs and allows the ranges of the most active molecular structures to be identified.

CARSO requires a reduced number of descriptors to be chosen according to the information coming from the linear PLS analysis. Accordingly, the descriptor variables 1 (MR for the R_1 substituent), 9 (t_1 for the R_8 substituent), and 3 (MW for the R_7 substituent) were chosen. The CARSO procedure works thereafter by expanding the descriptor matrix and building a new data matrix containing nine explanatory variables (x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2 , x_1x_3 , x_2x_3) and the dependent variable (D_{tot}).

The linear PLS analysis of the expanded matrix gives a three-component model that can be transformed, according to the described CARSO procedure, into a quadratic polynomial that represents the response surface.

The R^2 value of the PLS model on the expanded matrix is 0.65, but this relatively low value is not surprising since CARSO is not aimed at finding a better fit. This may be obtained by different techniques, e.g. polynomial fitting. On the contrary, CARSO, as stated in the last paragraph of the Methods section, uses an “a priori” chosen quadric equation in order to find the ranges of the descriptor variables between which the value assumed by the biological response within the experimental domain remains above a certain level. The response surface can be analytically studied by means of canonical analysis that allows the only existing stationary point (maximum or minimum or saddle point) within the domain to be found. In the CARSO procedure, when the response surface is not “bell-shaped”, the search for the constrained maximum within the experimental domain is made by means of the Lagrange analysis and seeks the values assumed by the response at the extreme points located at the borders of the domain. Our data set gave a saddle point (see Figure

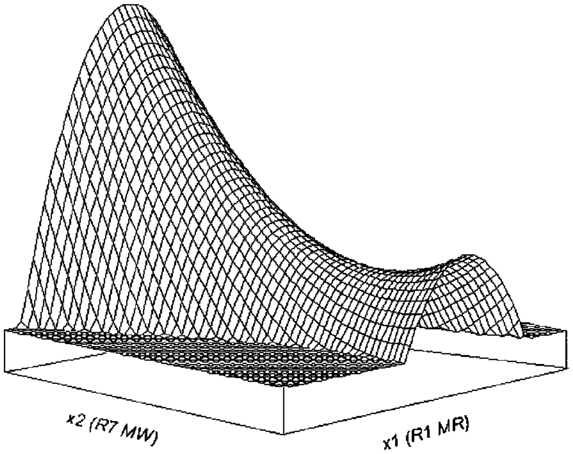


Figure 10. Three-dimensional response surface obtained by CARSO procedure.

Table 8. Results of the CARSO Analysis

10.9 < **X1 MR (R_1)** < 17.0
132 < **X2 MW (R_7)** < 161
1.18 < **X3 t_1 (R_8)**

10) and the resulting ranges for the best extreme point, corresponding to a y value of $D_{tot} = 0.79$, are reported in Table 8.

These ranges are referred to as structural descriptors and can be used to design new candidates to be synthesized and tested. In fact, the chemical information contained in the response surface model indicates that a cyclopropyl group is optimal in N-1 position, but other branched alkyl groups could also be good, and the groups containing two fused rings seem to be the most suitable at C-7. In fact, in the range for the R_7 only tetrahydroisoquinoline is found.

Unfortunately, the derived information for the R_8 is not so clear. The methyl group is a really good one, but this is just a qualitative and not a quantitative result. In fact, because of the poorly designed training set, which gathered the historical data produced by our group over a number of years, information is lacking about the activity of molecules with a larger group. Molecules with a methyl group are the best ones, and the methyl group is the largest substituent used in this position. Consequently, the model (Table 8) indicates an open range, and a larger substituent should provide a further increase of the antibacterial activity. However, we are aware that there should be a size limit for the R_8 substituent, after which it becomes too large to be accommodated into the receptor site. It is not known whether this limit has already been reached by the methyl group or if a larger group can be allocated.

Testing the Suggested Compounds. According to the interpretation of the quadratic model discussed in the previous section, we decided to synthesize and test the most promising structures, i.e. those bearing a larger substituent at C-8. Consequently, the molecules with an ethyl group and a methoxy group sitting at C-8 were prepared. For the former tetrahydroisoquinoline was the C-7 substituent, while for the latter, the C-7 group was *N*-methylpiperazine. For both molecules the N-1 substituents was cyclopropyl.

Although details of the syntheses and biological activities will be reported elsewhere,[16] none of the compounds exhibited an improved activity. On the contrary, while the methoxy derivative was only slightly less active than the 8-methyl congener, the activity of the 8-ethyl derivative against both Gram-positive and Gram-negative microorganisms dropped significantly.

The results show that the maximum volume allowed for the substituent at C-8 is sufficient to accept the methyl group, but not large enough to allocate even slightly larger substituents.

Conclusions

We were prompted to publish our work because of the appearance of the recent paper by Llorente et al.,[17] which is somewhat related to our study. This paper is based on "3D-QSAR" models derived by the APEX program,[18] which, even if some of the parameters used to describe molecules are intrinsically three-dimensional, is not a true 3D-QSAR tool, since it does not work on fields or energies derived at the nodes of a three-dimensional grid.

The main drawback is that APEX uses multivariate linear regression analysis (MLR) as the chemometric tool: Llorente at al. have a set of 15 structural descriptors, and under these conditions, because of the unavoidable multicollinearity, MLR is inappropriate, giving misleading results, and should be replaced by PLS. However, since the APEX program does not rely on appropriate cross-validation criteria,[8b,19] the derived models are apparently good, even when they are, in fact, nonpredictive. We are convinced that the Palumbo model[20] is preferable to the Shen model[21] for explaining the action of quinolones, but this support is based on the reasons expressed in the original paper, and not because of the apparent support given by Llorente et al.

The chemometric guidelines in the present work appear to be more reliable and show that, by properly applying chemometric strategies and tools, information can be extracted from a large amount of biological data, the existing data can be interpreted, and new structures designed. In other words, relatively simple, but carefully thought out, statistical analysis of a set of biological data can yield a wealth of information.[22]

Moreover, the current paper illustrates the value of graphical representation of statistical data as the relative activities, Gram-negative or Gram-positive specificities, and structural feature information has been combined into a set of graphical scatter plots of the data, yielding valuable information which might possibly have been missed without such a representation.[22]

In particular PCA has shown the redundancy of the available information (three groups of strains) and helps to indicate which structural feature affects each type of strain the most, while PLS permits the best compromise between the available activities to be described (expressed as total desirability) in terms of traditional, statistical, or topological descriptors.

Finally, the results of a RS study can be used to estimate the ranges within which it is possible to vary substituents at each site and indicate possible guidelines for hopefully increasing the overall activity. The fact that the activities of the suggested molecules were lower than those predicted, even though disappointing, should not raise doubts about the value of the chemometric approach or about the suitability of structural descriptors. On the contrary, it is reasonable the failure is mainly due to a poorly designed set of compounds, since the study undertaken did not follow strict design criteria. Such criteria were used in order to extract less unbalanced subsets from the available structures, but as stated earlier, we knew that all information necessary for making sound and reliable predictions outside the explored domain had not been collected. Consequently, this work once again demonstrates the need for a properly designed selection of informative structure for any QSAR study.

Acknowledgment. We thank the Italian funding agencies of MURST and CNR for financial support to A.F. and S.C. and Mediolanum Farmaceutici SpA for scientific collaboration.

References

(1) Bonelli, D.; Cecchetti, V.; Clementi, S.; Cruciani, G.; Fravolini, A.; Savino, A. F. The Antibacterial Activity of Quinolones against E. Coli: a Chemometric Study. *Quant. Struct.-Act. Relat.* **1990**, *10*, 333−343.

(2) Bonelli, D.; Cecchetti, V.; Clementi, S.; Fravolini, A.; Savino, A. F. Chemometric Rationalization of the Structural Features Affecting the Antibacterial Activity of Quinolones Against *Staphylococcus aureus*. *Pharm. Pharmacol. Lett.* **1993**, *3*, 13−16.

(3) Cecchetti, V.; Clementi, S.; Cruciani, G.; Fravolini, A.; Pagella, P. G.; Savino, A. F.; Tabarrini, O. 6-Aminoquinolones: A New Class of Quinolone Antibacterial? *J. Med. Chem.* **1995**, *38*, 973−982.

(4) Cecchetti, V.; Fravolini, A.; Lorenzini, M. C.; Tabarrini, O.; Terni, P.; Xin, T. Studies on 6-Aminoquinolones: Synthesis and Antibacterial Evaluation of 6-Amino-8-methylquinolones. *J. Med. Chem.* **1996**, *39*, 436−445.

(5) (a) Cecchetti, V.; Fravolini, A.; Fringuelli, R.; Mascellani, G.; Pagella, P.; Palmioli, M.; Segre, G.; Terni, P. Quinolinecarboxylic Acids. 2. Synthesis and Antibacterial Evaluation of 7-Oxo-2,3-dihydro-7H-pyrido[1,2,3-*de*][1,4]benzothiazine-6-carboxylic Acids. *J. Med. Chem.* **1987**, *30*, 465−473. (b) Mediolanum Rufloxacin Hydrochloride. *Drugs Future* **1991**, *16*, 678−680.

(6) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. *Chemometrics, Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; Reidel: Dordrecht, 1984; p 17.

(7) (a) Wold, H. Non Linear Estimation by Iterative Least Squares Procedures. In *Research Papers in Statistics: Festschrift for J. Neyman*; Wiley: New York, 1966. (b) Wold, S.; Johansson, E.; Cocchi, M. PLS - Partial Least-Squares Projections to Latent Structures. In *3D QSAR in Drug Design: Theory Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 523−550.

(8) (a) Clementi, S.; Cruciani, G.; Curti G. Some Application of the Partial Least-Squares Method. *Anal. Chim. Acta* **1986**, *191*, 149−160. (b) Clementi, S.; Wold, S. How to Choose the Proper Statistical Method. In *Chemometric Methods in Molecular Design;* van de Waterbeemd, H., Ed.; Vol. 2 of *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1995; pp 319−338.

(9) Wold, S.; Sjöström, M. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics Theory and Application*; Kowalski, B. R., Ed.; ACS Symposium Series: Washington, 1977; pp 243−282.

(10) SIMCA-S for Windows Manual; UMETRI AB: Umeå, 1994.

(11) Clementi, S.; Cruciani, G.; Curti, G.; Skagerberg, B. PLS Response Surface Optimization: the CARSO Procedure. *J. Chemom.* **1989**, *3*, 499−509.

(12) Bertuccioli, M.; Clementi, S.; Cruciani, G.; Giulietti, G.; Rosi, I. Food Quality Optimization. *Food Qual. Preference* **1990**, *2*, 1−12.

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

(14) Skagerberg, B.; Bonelli, D.; Clementi, S.; Cruciani, G.; Ebert, C. Principal Properties of Aromatic Substituents. A Multivariate Approach for Design in QSAR. *Quant. Struct.-Act. Relat.* **1989**, *8*, 32−38.

(15) Dunn, W. J., III; Wold, S. Structure-Activity Analyzed by Pattern Recognition: the Asymmetric Case. *J. Med. Chem.* **1980**, *23*, 595−599.

(16) Cecchetti et al., work in preparation.

(17) Llorente, B.; Leclerc, F.; Cedergren, R. Using SAR and QSAR Analysis to Model the Activity and Structure of the Quinolone-DNA Complex. *Bioorg. Med. Chem.* **1996**, *4*, 61−71.

(18) APEX-3-D, User Guide, Version 1.4, Biosym Technologies, 1993

(19) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191−193.

(20) Palù, G.; Valisena, G.; Ciarrocchi, G.; Gatto, B.; Palumbo, M. Quinolone binding to DNA is mediated by magnesium ions. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 9671−9675.

(21) Shen, L. L.; Baranowski, J.; Pernet, A. G. Mechanism of Inhibition of DNA Gyrase by Quinolone Antibacterial: Specificity and Cooperativity of Drug Binding to DNA. *Biochemistry* **1989**, *28*, 3879−3885.

(22) The authors thank one of the reviewers for these assessments

JM960385P