Antimicrobial Activity Characterization in a Heterogeneous Group of Compounds

R. García-Domenech* and J. V. de Julián-Ortiz†

Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain

Received November 20, 1997

In this work we carry out a study of pattern recognition to detect the microbiological activity in a group of heterogeneous compounds. The structural descriptors utilized are the topological connectivity indexes. The methods followed are stepwise linear discriminant analysis (linear analysis) and artificial neural network (nonlinear analysis). Although both methods are appropriate to differentiate between active and inactive compounds, the artificial neural network is, in this case, more adequate, since it shows in a test set a prediction success of 98%, versus 92% obtained with linear discriminant analysis.

INTRODUCTION

Different connectivity indexes (CIs) have shown their success in QSAR analysis to predict diverse physical, chemical, and biological properties in several groups of compounds.[1−6] An advantage of the CI is the swiftness of its calculation for almost any conceivable organic molecule. Because of this fact, it is easy to obtain such indexes for a huge data base of compounds. The Kier−Hall indexes are the most widely used CIs, and, although they have not at present an unambiguous interpretation, recent theoretical articles relate them with orbital energies.[7,8]

SAR also can be regarded as a problem of pattern recognition. Such techniques have been already applied to examine structural features that underlie patterns associated with biological effects.[9] Among these methods, we can point out the stepwise linear discriminant analysis (LDA) and the artificial neural networks (ANNs).

LDA finds linear combinations of variables that are able to distinguish between two or more groups. By means of structural descriptors, we can classify the compounds into active or inactive ones. The use of LDA has been relatively scarce in drug design, although it is a useful approach. For example, in recent work, our group has shown the ability of LDA as an excellent tool for drug design.[10−12]

Recently, artificial neural networks (ANNs) have been the center of attention in the field of pattern recognition. The ANN is a computer-based model in which a number of nodes, also called processing elements, units, or neurons, are interconnected by links in a netlike structure forming "layers"[13,14] (Figure 1). A variable value is assigned to every node. The nodes can be one of three different kinds:

(i) Input nodes, which receive their values by direct assignation and are associated with independent variables, with the exception of the bias node. They form the input layer.

(ii) Hidden nodes, which collect values from other nodes, giving a result that is passed to a noninput node. They constitute hidden layers.

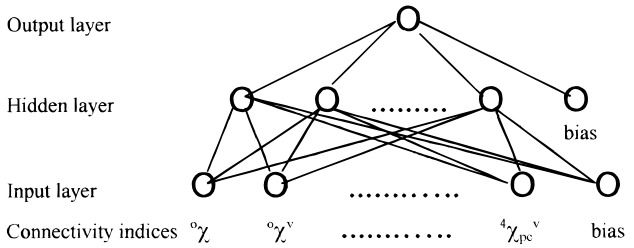

Figure 1. Topology of the neural network used in training set to predict the microbiological activity.

(iii) Output nodes, which collect values from other units. They correspond to different dependent variables, forming the output layer.

The links between units have values associated, named weights, that condition the values assigned to the nodes. There exist additional weights assigned to bias values that act as node value offsets. The weights are adjusted through a training process.

The characteristics of the ANN have been found to be suitable for data processing, in which the functional relationship between the input and the output is not previously defined. Another prominent characteristic of the ANN is its ability to classify or grade.

In this work we first present the results obtained by applying these two techniques, with CIs as descriptor variables, to a group of structurally heterogeneous antimicrobial compounds, and then we compare these results, and finally we examine the validity of the models obtained by searching for active compounds in other, more numerous groups.

METHODS, RESULTS, AND DISCUSSION

We have applied molecular connectivity to classify the microbiological activity of compounds using linear analysis, LDA, and nonlinear analysis, ANN. The topological descriptors used were the 16 Kier−Hall nonvalence and valence connectivity indexes up to fourth order,[15,16] excluding the chain terms.

We have used 111 compounds, divided into two groups: a training set group with 64 compounds (of which 34 show

* E-mail: ramon.garcia@uv.es.
† E-mail: julian@colom.combios.es.

S0095-2338(97)00245-X CCC: $15.00
Published on Web 04/15/1998

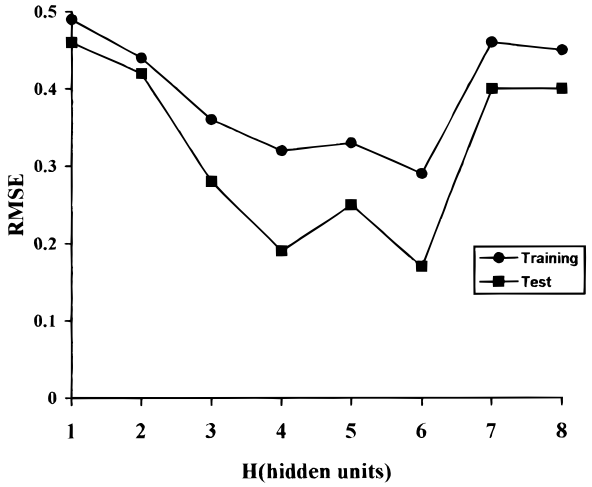

Figure 2. RMSE of training and test sets as a function of the number of hidden units.

contrasted microbiological activity) and a test set group with 47 compounds (of which 26 are actives). The assignment of the activity of each compound follows the Merck Index nomenclature.

The method used for descriptor selection with linear discriminant analysis (LDA) was the *F*-Snedecor parameter. The classification criterion was the minimum value of Mahalanobis.[17] The quality of the discriminant function is evaluated through Wilk's *U*-statistical parameter. LDA was performed using module 7M of the package program BMDP.[18]

The discriminant function selected was

$$Y = -5.21({}^{0}\chi^{v}) + 11.77({}^{2}\chi) - 16.65({}^{3}\chi_{c}) + 13.56({}^{3}\chi_{c}{}^{v}) + 186.35({}^{4}\chi_{c}) - 0.56({}^{4}\chi_{pc}) - 3.24({}^{4}\chi_{pc}{}^{v}) - 3.40$$

$$N = 111 \qquad F = 20.9 \qquad U \text{ statistic} = 0.2762$$

The topology of the neural network used is shown in Figure 1. Input variables were not normalized for the sake of simplicity. The number of inputs was 16, and a unique output node was used. In the training set, the active compounds were assigned an output value of 1, while the inactives had a 0 value. The optimal number of hidden units is the one that minimizes the root-mean-square error (RMSE) in a test set, not used in net training. In Figure 2, it can be seen that the training set and test set RMSE is a concave curve, with a minimum for 6 hidden units.

Neural network training was performed through the back-propagation algorithm with the momentum method and weight decay. The natural sigmoid was used as activation function.[13] The training stop was based on minimization of the RMSE in the test group. The learning rate, threshold value, momentum, and weight decay values were set to be 0.1, 0, 0.1, and 0.0001, respectively. The values of initial weights were randomly assigned in a range of ±0.1. The computer program used was NevProp.[19]

Tables 1−3 show the results obtained applying both methods. Those compounds that showed a probability or output in the range 0.6−1.0, were classified as active (+) and they were classified as inactive (−) if the probability or output was between 0 and 0.4; they were nondefined (±) in other cases. Overfitting is avoided through the validation in a test set.

LDA shows better predictions than ANN in the training set (94% correct prediction for LDA versus 89% for ANN). Nevertheless, applying both methods to the test set, the ANN gives a microbiological activity prediction superior to that obtained by LDA (98% of correct prediction by ANN versus 92% by LDA).

All the sets of compounds employed were structurally heterogeneous. It can be seen in Table 1 that, in the composition of the training actives set, there are only quinolone analogues, sulfonamides, and cephalosporins. There is no apparent structural reason why four compounds do not fit the clustering predictions: nalidixic acid is a quinolone analogue, cefroxadine is a cephalosporin, and the others are sulfonamides. The test actives set has antibacterial compounds from groups not included in the training set: trimethoprim (a diaminopyrimidine), cefbuperazone and cefotetan (two cephamicins), moxalactam (an oxacephem), and dapsone (a sulfone). Of these, only trimethoprim was misclassified. It is clear that there exist a topologically common pattern more subtle than classifications attending to functional groups.

While in linear regression models it is desirable to have at least five data points for each adjustable parameter or coefficient in the discriminant function, no corresponding rule of thumb exists for ANN models. The number of bonds *P* (number of adjustable parameters) is calculated in this study by

$$P = (I + 1)H + (H + 1)O$$

where *I*, *H*, and *O* are the number of input, hidden, and output units, respectively. In modeling methods that rely on adjustable parameters, the following ratio is defined:

$$\rho = (\text{no. of data points})/(\text{no. of adjustable parameters})$$

In studies with ANN, it is recommended[20] to have $\rho > 2$, although other criteria can be found in the literature.[21] In linear models, $\rho > 4$. The reason why in nonlinear models the criterion is more relaxed is not clear to us. We think that the predictive power of test groups is similar to that of the training group; except for the chance effects, this model is valid. Among several valid models, we choose the one with a lower ρ, because it will be the simplest one. The cross-validation procedure also solves the question raised about fortuitous correlations when too many variables are screened.[22]

In our ANN, no. of data points = 64, no. of adjustable parameters = $P = 17 \times 6 + 7 = 109$. Clearly, $\rho < 1$, which is not an obstacle to obtaining excellent prediction in the test group. It seems that our model, with one hidden layer and $H = 6$, extracted the features without memorizing the data set. Maybe this ρ value reveals redundancies in input information that could be avoided with an adequate choice of descriptors, a topic beyond the scope of the present work.

In both models, the prediction reached is high enough that is is unnecessary to introduce additional descriptors that could afford more information. The results obtained in this study justify the use of simple descriptors alone; also, they make

Table 1. Results Obtained by LDA and ANN on the Training Group

active compounds					inactive compounds				
	LDA		ANN			LDA		ANN	
compound	prob.	class	output	class	compound	prob.	class	output	class
cefaclor	0.771	+	0.898	+	alclofenac	0.001	−	0.042	−
cefamandole	0.988	+	0.899	+	aminopyrine	0.001	−	0.040	−
cefatrizine	0.999	+	0.722	+	ampyrone	0.002	−	0.040	−
cefazedone	0.945	+	0.899	+	artromialgina	0.005	−	0.889	+
cefazolin	1.000	+	0.834	+	azapropazone	0.226	−	0.042	−
cefoperazone	1.000	+	0.893	+	benoxaprofen	0.012	−	0.074	−
cefoxidine	1.000	+	0.892	+	butibulen	0.007	−	0.045	−
ceipiramide	0.999	+	0.899	+	diclofenac	0.000	−	0.040	−
cefroxadine	0.985	+	0.400	−	epyrizol	0.002	−	0.047	−
ceftizoxime	0.942	+	0.898	+	etodolac	0.183	−	0.041	−
cephalexin	0.886	+	0.899	+	fenacetine	0.001	−	0.041	−
cephaloglycin	0.975	+	0.899	+	fenodol	0.004	−	0.051	−
cinoxacin	0.911	+	0.899	+	fenoprofen	0.002	−	0.051	−
enoxacin	0.992	+	0.899	+	flutenamic acid	0.003	−	0.898	+
enrofloxacin	1.000	+	0.891	+	ibuprofen	0.012	−	0.054	−
flumequine	0.898	+	0.744	+	indomethacin	0.001	−	0.051	−
lomefloxacin	0.996	+	0.865	+	methopholine	0.005	−	0.040	−
nalidixic acid	0.005	−	0.899	+	naproxen	0.001	−	0.042	−
norfloxacin	0.985	+	0.895	+	paracetamol	0.000	−	0.087	−
ofloxacin	1.000	+	0.899	+	perisoxal	0.993	+	0.882	+
oxolinic acid	0.858	+	0.866	+	phenopyrazone	0.045	±	0.044	−
pefloxacin	0.971	+	0.899	+	phenylbutazone	0.016	−	0.041	−
piperamic acid	0.976	+	0.899	+	phenylsalicylat	0.003	−	0.063	−
piromidic acid	0.897	+	0.898	+	piperilone	0.592	±	0.042	−
sulfadiazine	0.998	+	0.891	+	piroxicam	0.006	−	0.092	−
sulfadimethoxine	0.995	+	0.260	−	salicylic acid	0.007	−	0.173	−
sulfadoxine	0.991	+	0.899	+	sulindac	0.012	−	0.792	+
sulfamerazine	0.997	+	0.876	+	tolmetin	0.002	−	0.124	−
sulfamethazine	0.996	+	0.899	+	viminol	0.000	−	0.040	−
sulfamethoxazole	1.000	+	0.889	+	zomepyrac	0.000	−	0.043	−
sulfamethoxidiazine	0.997	+	0.898	+					
sulfathiazole	0.998	+	0.715	+					
sultdurazin	0.997	+	0.899	+					
sulfisoxazole	0.998	+	0.557	±					

Table 2. Results of Cross-Validation Analysis Obtained by LDA and ANN

active compounds					inactive compounds				
	LDA		ANN			LDA		ANN	
compound	prob.	class	output	class	compound	prob.	class	output	class
cefadroxil	0.977	+	0.899	+	acetanilide	0.001	−	0.040	−
cefazaflur	1.000	+	0.899	+	aminopropylon	0.001	−	0.040	−
cefazone	1.000	+	0.899	+	bumadizon	0.005	−	0.343	−
cefbuperazone	0.857	+	0.899	+	carprofen	0.017	−	0.042	−
celmenoxime	0.975	+	0.899	+	cinmetacin	0.001	−	0.079	−
celmetazole	0.001	+	0.751	+	clidanac	0.919	+	0.043	−
cefonicid	1.000	+	0.899	+	clopirac	0.001	−	0.041	−
ceforanide	0.998	+	0.899	+	chlorthenoxazin	0.051	−	0.041	−
cefotaxime	0.061	−	0.899	+	difenpiramide	0.001	−	0.041	−
cefotetan	1.000	+	0.899	+	ethenzamide	0.001	−	0.041	−
cefotiam	1.000	+	0.899	+	ethoheptazine	1.000	+	0.040	−
cefsulodin	1.000	+	0.899	+	fentiazac	0.001	−	0.041	−
ceftazidime	1.000	+	0.899	+	feprazone	0.008	−	0.042	−
ceftezole	0.999	+	0.899	+	iboproxam	0.016	−	0.120	−
ceftriaxone	0.985	+	0.899	+	isopyrin	0.005	−	0.040	−
cefuroxime	0.997	+	0.899	+	kebuzone	0.025	−	0.068	−
cephacetrile	0.882	+	0.899	+	morazone	0.271	−	0.040	−
cephaloridine	0.967	+	0.899	+	oxametacine	0.000	−	0.098	−
cephalothin	0.981	+	0.899	+	oxyzincofen	0.009	−	0.152	−
cephapirin	0.935	+	0.899	+	salsalate	0.010	−	0.899	+
cephradine	0.972	+	0.899	+	tiaprofenic acid	0.001	−	0.041	−
ciprofloxacin	1.000	+	0.899	+					
dapsone	1.000	+	0.899	+					
fleroxacin	1.000	+	0.899	+					
moxalactam	1.000	+	0.899	+					
trimethoprim	0.001	−	0.740	+					

it unnecessary to consider more complex architectures, such as additional layers of hidden units.

In the linear equation, the most significant variables are $^{4}\chi_{c}$, $^{0}\chi^{v}$, $^{3}\chi_{c}$, and $^{2}\chi$. The corresponding *F* values to select every variable are 33.98, 23.66, 14.88, and 14.10, respectively. In the ANN, the most weighted nodes correspond, in this order, to the descriptors $^{4}\chi_{c}^{v}$, $^{4}\chi_{c}$, $^{3}\chi_{c}^{v}$, $^{3}\chi_{c}$, and $^{4}\chi_{pc}^{v}$. The quantification of the weight taken is the sum—in absolute values—of the weights that connect the corresponding input node with the hidden layer, divided by the medium value of

Table 3. Comparison of Artificial Neural Networks and Linear Discriminant Analysis for Predicting Microbiological activity[a]

	no. of data points		training set		test set	
method	training	test	no. of outliers	% success	no. of outliers	% success
LDA	64	47	2	93.8	5	91.5
ANN	64	47	6	89.0	1	97.9

[a] Results classified as ± are not counted as outliers, nor do they contribute to % success.

the index in the training group. Thus, for the five indexes above mentioned, the following estimates of relative weight were obtained: 452.32, 427.11, 16.50, 10.32, and 9.51. It is noteworthy that $^4\chi_c$ and $^3\chi_c$ are important in both models. This fact demonstrates consistency between the methods, and also the determinant influence of the existence of cluster-type subgraphs in the molecular graph. In spite of this, it is not possible to establish a physical interpretation of the results, and by no means can we develop a mechanistic model, keeping in mind that the mechanisms of action of the antimicrobials are not considered to include them in the training set. QSAR analysis is not necessarily a causal analysis, but rather is the search of phenomenological correlations that fit and are adequate to search new active compounds. The extraction of mechanistic conclusions from other QSAR analyses, such as the Hansch approach, has been criticized,[23] although these kinds of descriptors have a priori precise meanings and easy interpretation.

As can be seen in Tables 1 and 2, there are no false positives given by the two methods simultaneously. To find new compounds with antimicrobial activity, the two models were used to screen a data base of structurally diverse compounds containing the 16 descriptors calculated for each one. Several compounds were found that showed activity values for both functions. Among these are 1-(4-nitrophenyl)piperazine and ethylenediaminetetraacetic acid. The corresponding values of *Y* were 4.32 (probability of activity, 0.998) and 4.95 (probability of activity, 0.992), respectively. The outputs of the ANN were, respectively, 0.882 and 0.899. The antimicrobial activity of these two chemicals had been identified in previous works,[24−28] which confirms the validity of the method to predict such activity in further sets of compounds.

CONCLUSION

Both methods afford a powerful discriminant tool when using connectivity indexes, beyond chance effects. This is revealed through the test group results, which are comparable with the training group statistics.

ANN is an interesting approach to enhance a descriptive model; on the other hand, the longer computational times required for developing a neural network relative to linear regressions can be somewhat inconvenient but is not a limiting feature.

This work proposes valid models to predict the antimicrobial activity of compounds with heterogeneous structures. These models are based on topological patterns that skip over the parent skeleton concept, which is the basis of the congeneric series notion.

ACKNOWLEDGMENTS

We thank Professor Ramon Carbó (Universitat de Girona) and Professor Jorge Gálvez for careful reading of the manuscript and for valuable comments. We are also indebted to Jorge Gálvez and Javier Llombart for the development of the INDIS programs for χ calculations and to Carolina de Gregorio for her worthy advice. This study has been supported by CICYT, SAF96-0158-C02-02 (Ministerio Español de Educación y Cultura), and GV-3300/95 (Generalitat Valenciana). Also, we acknowledge the positive suggestions given by the reviewers for the improvement of the present paper.

REFERENCES AND NOTES

(1) Kier, L. B.; Hall, L. H. The Nature of Structure-Activity Relationships and their Relation to Molecular Connectivity. *Eur. J. Med. Chem.-Chim. Ther.* **1977**, *12*, 307−312.
(2) Kier, L. B.; Hall, L. H. The Relation of Molecular Connectivity to Molecular Volume and Biological Activity. *Eur. J. Med. Chem.-Chim. Ther.* **1981**, *16*, 399−407.
(3) García-Domenech, R.; Gálvez, J.; Moliner, R.; García-March, F. Prediction and Interpretation of Some Pharmacological Properties of Cephalosporins Using Molecular Connectivity. *Drug Invest.* **1991**, *3*, 344−350.
(4) Julián-Ortiz, J. V. de; García-Domenech, R.; Gálvez, J.; Soler, R.; García-March, F.; Antón-Fos, G. M. Use of Topological Descriptors in Chromatographic Chiral Separations. *J. Chromatogr. A* **1996**, *719*, 37−44.
(5) Antón-Fos, G. M.; García-Domenech, R.; Pérez-Giménez F.; Peris-Ribera, J. E.; García-March, F.; Salabert-Salvador, M. T. Phamacological Studies of the Two New Hypoglycaemic Compounds 4-(3-Methyl-5-oxo-2-pyrazolin-1-yl)benzoic Acid and 1-(Mesitylen-2-sulfonyl)-1H-1,2,4-triazole. *Arzneim-Forsch./DrugRes.* **1994**, *44*, 821−826.
(6) Gálvez, J.; García-Domenech, R.; Julián-Ortiz, J. V. de; Soler, R. Topological Approach to Analgesia. *J. Chem. Inf: Comput. Sci.* **1994**, *34*, 1198−1203.
(7) Stankevich, I. V.; Skovortsova, M. I.; Zefirov, N. S. On a quantum chemical interpretation of molecular connectivity indices for conjugated hydrocarbons. *J. Mol. Struct. (THEOCHEM)* **1995**, *342*, 173−179.
(8) Gálvez, J. On a topological interpretation of electronic and vibrational molecular energies. J. *Mol. Struct. (THEOCHEM)*, in press.
(9) Aoyama, T.; Suzuki Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(10) Gálvez, J.; García-Domenech, R.; Julián-Ortiz, J. V. de; Soler R. Topological Approach to Drug Design *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 272−284.
(11) Gálvez, J.; Gómez-Lechón, M. J.; García-Domenech, R.; Castell, J. V. New Cytostatic Agents Obtained by Molecular Topology. *Bioorg. Med. Chem. Lett.* **1996**, *19*, 2301−2306.
(12) García-Domenech, R.; Gregorio Alapont, C. de; Julián-Ortiz, J. V. de; Gálvez, J.; Popa, L. Molecular Connectivity to Find β-Blockers with Low Toxicity. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 567−572.
(13) Sumpter, B. G.; Getino, C.; Noid, D. W. Theory and Applications of Neural Computing in Chemical Science. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439−481.
(14) Aoyama, T.; Suzuki Y.; Ichikawa, H. Neural Networks Applied to Structure-Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905−908.
(15) Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity Series Concept Applied to Density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230.
(16) Kier, L. B.; Hall, L. H. General Definition of Valence Delta-Values for Molecular Connectivity. *J. Pharm. Sci.* **1983**, *72*, 1170−1173.
(17) Costanza, M. C.; Afifi, A. A. Comparation of Stopping Rules in Forward Stepwise Discriminant Analysis. *J. Am. Stat. Assoc.* **1979**, *74*, 777−785.
(18) Dixon, W. J. BMDP Statistical Software, University of California, Berkeley, 1990.
(19) Goodman, P. H. NevProp Neural Network Simulator, University of Nevada, Reno, 1995.
(20) Manallack, D. T.; Livingstone D. J. Artificial Neural Networks: Application and Chance Efects for QSAR Data Analysis. *Med. Chem. Res.* **1992**, *2*, 181−190.
(21) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in

Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(22) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of QSAR. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(23) Verloop, A. Linear Free Energy Parameters in Drug Design. In *Drug Design. Vol. III*; Ariëns, E. J., Ed.; Academic Press: New York, 1972; pp 167−170.
(24) Adler, H. E.; DaMassa, A. J.; Scott, W. F. Studies on egg disinfection. *Poult. Sci.* **1979**, *58*, 799−806.
(25) Seeger, K.; Hentschel, G. Effect of ethylenediaminetetraacetic acid on the antibacterial activity of some antibiotics and their enteral absorption in the pig. *Arzneim-Forsch./Drug Res.* **1971**, *21*, 1590−1594.
(26) Chew, B. P.; Tjoelker, L. W.; Tanaka, T. S. In vitro growth inhibition of mastitis causing bacteria by phenolics and metal chelators. *J. Dairy. Sci.* **1985**, *68*, 3037−3046.
(27) Kida, N.; Suzuki, S.; Yamanaka, T.; Furoyama, K.; Taguchi, F. Effect of pH on preferential antibacterial-activity of ethylenediaminetetraacetic acid (EDTA). *Nippon Saikingaku Zasshi* **1992**, *47*, 625−629.
(28) Gálvez, J.; García-Domenech, R.; Gregorio Alapont, C. de; Julián-Ortiz, J. V. de; Salabert-Salvador, M. T.; Soler-Roca, R. New antibacterial drugs designed by Molecular Connectivity. In *Advances in Molecular Similarity. Vol. I*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press Inc.: London, 1996; pp 267−280.

CI9702454