CROATICA CHEMICA ACTA CCACAA **75** (2) 545–562 (2002)

ISSN-0011-1643
CCA-2816
Original Scientific Paper

Quantitative Structure-Activity Relationship of Tricyclic Carbapenems: Application of Artificial Intelligence Methods for Bioactivity Prediction*

Mira Lebez,[a] *Tom Šolmajer,*[b,c] *and Jure Zupan*[d,**]

[a] *Faculty of Chemistry and Chemical Technology, University of Ljubljana, Ljubljana, Slovenia*

[b] *Laboratory of Molecular Modeling and NMR Spectroscopy, National Institute of Chemistry, Hajdrihova 19, P. O. Box 660, 1001 Ljubljana, Slovenia*

[c] *Lek, d.d., Research and Development, Celovska 135, 1526 Ljubljana, Slovenia*

[d] *Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, P. O. Box 660, 1001 Ljubljana, Slovenia*

Received May 7, 2001; revised October 29, 2001; accepted October 30, 2001

Resistance to antibiotics in bacterial population has widened the interest of scientific community for development of novel therapeutic compounds. Penicillins and cephalosporins which share the β-lactam structural moiety form the most abundant group of antibiotics on the market. Their recently developed tricyclic analogues have shown remarkable bioactivity towards broad spectrum of bacterial species. In a series of 52 tricyclic carbapenems represented by the 180'dimensional »spectrum-like« representation we studied the structure-activity relationships by application of an artificial neural network. The molecular structure representation by spectral intensity values served as inputs into the counter-propagation artificial neural network (CP-ANN). SIMPLEX optimization was carried out to obtain the best ANN model and a genetic algorithm approach was subsequently used to simultaneously minimize the number of variables. Thus, a search for the substituents that predominantly influence the experimental bioactivity was performed.

[*] This paper is dedicated to Professor Milan Randić (Des Moines) in honor of his 70th birthday.

[**] Author to whom correspondence should be addressed. (E-mail: jure.zupan@ki.si)

The constructed CP-ANN model yielded bioactivity values predictions with a correlation coefficient of 0.88, with their values extended over 4 orders of magnitude. The list of substituents selected by our automatic procedure can be compared with the data obtained by protein crystallography of the β-lactam inhibitors in complex with D,D-peptidase enzyme.

Key words: QSAR, tricyclic carbapenem derivatives, antibiotic activity, articial neural networks, genetic algorithms.

INTRODUCTION

Resistance to antibiotics is currently one of the major issues in modern therapy of infectious disease.[1,2] Since the discovery of penicillin more than half a century ago the β-lactam structures have played the central role in the most important antibiotics on the market. The nonselective and abundant use of antibiotics has caused the spread of bacterial resistance to antibiotics of penicillin and cephalosporin core structural moieties. Recently discovered carbapenem β-lactams with additional 5,6 or 7-membered ring fused to the penicillin 5-membered ring as a scaffold which posess interesting and broad spectrum antibiotic activity has prompted us to apply artificial intelligence methodology to study quantiative structure-activity relationship in this series in order to get insight into which structural features of these analogues relate to their bioactivity.

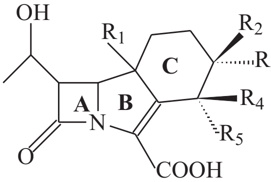

Figure 1. Schematic representation of a tricyclic carbapenem. The substituted ring C is fused to carbapenem nucleus A and B.

The goal of producing high-affinity ligands to target protein molecules requires the understanding of relationship between chemical structure and their properties called structure-activity relationships (SAR).[3] Several methods have been instrumental in explicit understanding of SARs, most notably molecular modeling[4,5] and comparative molecular field analysis (COMFA).[6] Alternatively, principles of artificial inteligence have been put to valuable use for such purpose.[7] In particular, artificial neural networks as source of a model for SAR coupled with the concept of evolutionary programming (genetic algorithms) as model optimisation technique has recently been put forward and used in a variety of examples such as spectroscopy applications,[8,9] combinatorial library optimisations for medicinal and organic chemistry[10] and SAR in various inhibitor-enzyme systems such as protein tyrosine kinase and thrombin.[11,12] In this work we extend the methodology with two novel algorithmic steps: (i) the spectrum like chemical structure representation was used to provide a uniform variable dimensionality and enhance the ease of variable manipulation by uniform length vector like description of molecules and (ii) SIMPLEX method was used to optimise the ANN model *i.e.* the selection of variables was optimized. The genetic algorithm approach was used in the final step (iii) to reduce the number of variables and obtain the correlation of substituent position on the molecular scaffold which was correlated with the bioactivity of the series of analogues. These steps are described in more detail in the Methods section below. The resulting ANN models and their variable reduction by the use of GA are given in the Results and Discussion section where we also compare the computed bioactivity values with their experimental counterparts for the test and control set of compounds which were previously not used in either training or testing procedures.

METHODS

The method used in the present work is described below and consists of the following steps:

Data Set Preparation

A complete search of the CAS Registry database[13,14] was first carried out to identify all tricyclic carbapenems with a measured antibiotic activity.[15–37] 74 compounds had data on structure and measured antibiotic activity of a total of 65 different bacterial strains was published. The antibiotic activity expressed as MIC (minimal inhibitory concentration) and measured for the bacterial strain *Clostridium perfringens* 615E was chosen as our data base since the bioactivity data were available for the largest number of, and structurally most diverse set of compounds. In Table I the biological and structural data on 52 compounds were collated.[15–34]

Structurally, the compounds belong to 12 different tricyclic skeletons represented in Figure 2 which have the ring C (see Figure 1) fused to the carbapenem nucleus and are substituted with substituents R_1-R_6. The fused ring can have a heteroatom substituted at various positions of this ring (rings C3-C10). In column 2 of Table II the value of index *i* represents the type of ring C as given in Figure 2

Chemical Structure Representation

3D structures of compounds can be alternatively represented by the recently introduced spectral representation.[38] Such representation is reversible, unique and uniform and due to its vectorial bit-like nature is particularly suitable for further

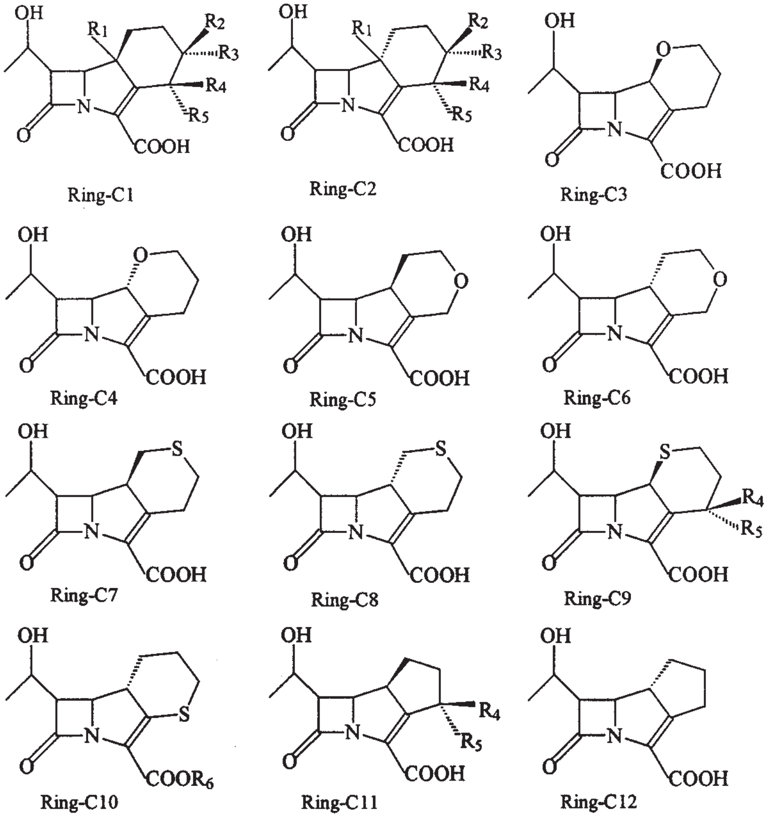

Figure 2. Schematic representation of 12 different rings C present in our data set of 52 tricyclic carbapenems.

use in artificial network models and genetic algorithms which require a uniform number of variables (descriptors) for each chemical structure. The structure of a molecule with N atoms in three dimensions where N can vary depending on the compound structure is transformed to a m-dimensional vector with components of variable intensity but of constant length m. The intensity is a function of position of a given atom relative to the chosen origin and charge on the atom. The compounds in our data set are particularly amenable for use of this representation since a large part of the molecule is identical in all compounds and the structure of the molecules in the set relatively rigid. This property is desirable since it simplifies the choice of the origin and eliminates the need for extensive and error-prone dihedral angle optimization procedures for each compound.

Compound Structure Optimization

A standard AM1 method as available in the computer program package Spartan[39] was utilized to compute the minimal energy geometry of all the molecules.

TABLE I

Antibiotic activity expressed as MIC values in strain *C. perfringens* 615E for the data set of 52 tricyclic carbapenems (Refs 15–34). In column 2 the value of index *i* represents the type of ring C as given in Figure 2.

No.	*i*	$\frac{\text{MIC}}{\mu\text{g mL}^{-1}}$	R$_1$	R$_2$	R$_3$	R$_4$	R$_5$	R$_6$
1	1	≤0.01	H	H	H	H	OCH$_2$CH$_2$F	H
2	1	≤0.01	H	H	H	H	OCH$_2$CH$_2$I	H
3	1	0.01	H	H	H	H	NHCH$_2$-C$_6$H$_5$	H
4	1	0.03	H	H	H	H	OCH$_3$	H
5	1	0.03	H	H	H	H	OCH$_3$	H
6	1	0.03	H	H	H	H	OCH$_2$CH$_2$OCH$_3$	H
7	1	0.03	H	H	H	H	OCH$_2$CH$_2$C≡N	H
8	1	0.03	H	H	H	H	OCH$_2$CH$_2$CH$_2$OH	H
9	1	0.03	H	H	H	H	OCH$_2$CH$_2$N=N$^+$=N$^-$	H
10	1	0.03	H	H	H	H	OCH$_2$CH$_2$N=N$^+$=N$^-$	H
11	1	0.03	H	H	H	H	NHCH$_2$-C$_6$H$_4$-*p*-NO$_2$	H
12	1	0.03	H	H	H	H	NHCH$_2$-C$_6$H$_4$-*p*-SO$_2$-NHCH$_2$CH=CH$_2$	H
13	1	0.03	H	H	H	H	NHCH(CH$_3$)-C$_6$H$_4$-*p*-NO$_2$	H
14	1	0.06	H	H	H	H	OCH$_2$CH$_2$OH	H
15	1	0.06	H	H	H	H	OCH$_2$CH$_2$CH$_2$NH$_2$	H
16	1	0.06	H	H	H	H	OCH$_2$CH$_2$CH$_2$N=N$^+$=N$^-$	H
17	1	0.06	H	H	H	H	NHCH$_2$-C$_6$H$_4$-*p*-F	H
18	1	0.10	H	H	H	H	H	H
19	1	0.10	H	H	H	H	N(CH$_3$)CH=NH	H
20	1	≤0.12	H	H	H	H	N(CH$_3$)C(=O)NH$_2$	H
21	1	≤0.12	H	H	H	H	N(CH$_3$)C(=O)NHC$_2$H$_5$	H
22	1	≤0.12	H	H	H	H	N(CH$_3$)C(=O)NH(*t*-Bu)	H
23	1	≤0.12	H	H	H	H	N(CH$_3$)C(=O)NHCH$_2$CH$_2$OH	H
24	1	0.12	H	H	H	OCH$_3$	H	H
25	1	0.12	H	H	H	H	OCH$_2$CH$_2$NH$_2$	H
26	1	0.12	H	H	H	H	OCH$_2$CH$_2$NHCH=NH	H
27	1	0.12	H	H	H	H	OCH$_2$CH$_2$NHC(=N$^+$H)CH$_3$	H
28	1	0.25	H	H	H	H	OCH$_2$CH$_2$N(CH$_3$)$_3$	H
29	1	0.50	H	H	OCH$_3$	H	H	H
30	1	0.50	H	H	H	H	NHCH=NCH$_3$	H
31	1	0.60	H	H	H	H	NHCH$_2$-C$_6$H$_3$(*m*-NO$_2$-*p*-Cl)	H

TABLE I (cont.)

No.	*i*	MIC / $\mu g\ mL^{-1}$	R_1	R_2	R_3	R_4	R_5	R_6
32	1	2.00	OCH_3	H	H	H	H	H
33	1	2.00	H	H	OCH_3	H	H	H
34	1	2.00	H	H	$CH_2CH_2CH_3$	H	H	H
35	1	4.00	H	H	CH_2CH_2OH	H	H	H
36	1	8.00	H	OCH_3	H	H	H	H
37	2	1.00	H	H	H	OCH_3	H	H
38	2	1.00	H	H	H	H	H	H
39	2	2.00	OCH_3	H	H	H	H	H
40	2	4.00	H	H	H	H	OCH_3	H
41	2	>32.00	H	H	H	H	$OCH_2CH_2OCH_3$	H
42	9	0.09	H	H	H	H	SCH_3	H
43	9	0.03	H	H	H	H	H	H
44	3	0.06	H	H	H	H	H	H
45	11	0.12	H	H	H	H	OCH_3	H
46	7	0.25	H	H	H	H	H	H
47	5	0.50	H	H	H	H	H	H
48	10	0.50	H	H	H	H	H	$CH_2CH=CH_2$
49	4	1.00	H	H	H	H	H	H
50	8	1.00	H	H	H	H	H	H
51	12	1.00	H	H	H	H	H	H
52	6	2.00	H	H	H	H	H	H

Mullikan charge distribution was computed in each molecule to supplement the spectral representation of compounds as described above and in Ref. 38. The width of lines in the spectral representation of compounds is dependent on atomic charges.

Application of an Artificial Neural Network Yields a Nonreduced QSAR Model

ANN Training Step

In house written CP-ANN and genetic algorithm packages[40] were used for these tasks.

The spectral values for the *m*-dimensional vector served as inputs for a counter-propagation artificial neural network[41] (CP-ANN). We have chosen $m = 180$ in order to compromise between the number of atoms in the compounds of the series ($N < 60$) and the number of variables used for correlation of bioactivity with structure.

This method was demonstrated previously to give better results than the similar back-propagation ANN if the number of items in the database is small as is the case with a small number of molecules $n = 52$ in our data set.

The leave-one-out cross-validation (LOO CV) test trained the network using the data for n-1 compounds and predicted the experimental antibiotic activity value for the n-th compound. This procedure was repeated n times yielding the list of n predictions. These predictions were ordered by difference between the experimental and computed bioactivity Δ. The first 22 compounds with largest absoute value of Δ were put in the training set taking into account the fact that the training set should contain as many compounds as possible to fully cover the entire information space and the remaining 30 compounds were sequentially stored in training, test and prediction set, respectively. In order to carry out the subsequent genetic algorithm (GA) procedure the compounds were thus divided into three groups: (i) a training set comprising 32 inhibitors, (ii) a test set of 10 inhibitors and (iii) a prediction set of 10 inhibitors. The initially removed prediction set provided a means to test the quality of the final model and did not influence either the determination of CP ANN model parameters like number of training epochs, number of neurons, maximal and minimal learning rates, *etc.*, or the next step of the procedure in which the selection of variables with GA is performed (step (e) below).[35]

The CP-ANN procedure was performed on the training set of 32 molecules and the antibiotic activity values in the test set of $N = 10$ compounds were predicted by the network as described below. During the training the weights of the winning neuron and close neighbours are corrected in small steps and thus in the last cycle they are completely adapted to the input object. By training the network we obtained a model whose prediction ability is determined by checking it with the test objects.

ANN Testing Step

The testing procedure is as follows: each test object X first finds (Figure 3) in the upper (Kohonen) layer the most similar neuron. The corresponding weight at the position j_x, j_y in the lower (output) layer gives the predicted target value.

The difference between target T and the predicted property P is squared and summed over all objects of the test set to give the PRESS (predicted residual errors sum of squares) measure of error. The regression equation linking theoretical and predicted properties is also determined. The correlation coefficient r derived from this equation represents the quality measure of the proposed model.

Representation Reduction by Use of GA Approach

In the last step of the present algorithm the complete representation of all compounds in the data set was reduced to its most relevant part by the use of a genetic algorithm.

A genetic algorithm consists of three basic processes mimicking Darwinian evolution: crossover, mutation and selection. In the crossover step, new chromosomes are generated by mixing those of the parents. In the mutation step, individual bits of the chromosome are exchanged at random and, in the selection step the best chromosomes are identified for the next round.

This last step should yield the lowest possible number of bits in the compound vector representation which gives a satisfactory prediction for biological activity for

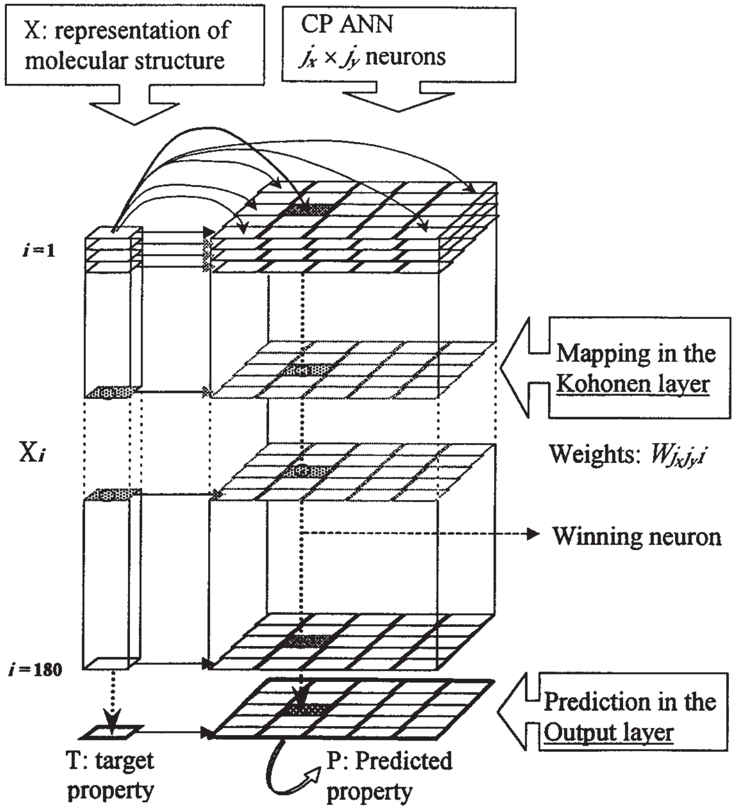

Figure 3. Schematic representation of a counter-propagation artificial neural network (CP ANN) used for predicting biological property (target T) from molecular structure (X) represented with 180 values of »spectrum like« intensity as described in the text above. During the training the weights (*W*) of the winning neuron (the winning neuron is determined with the position (j_x, j_y) in the Kohonen layer according to the similarity between the object (X_i, i = 1…180) and the neuron's weights (Wj_xj_yi, i = 1…180)) are adapted step by step, in learning cycles, to the compounds of the object.

each individual compound in the data set. The number of input parameters (180) representing the spectrum like code determines the length of the chromosome in bits. In our specific example the 180 components which represent each molecule in its spectrum-like representation were reduced to a few which provide the maximal influence on the experimentally determinded antibiotic activity of the compound.

The following selection procedure was implemented: each bit in the allele could take a value of 0 or 1. A value of 1 was used to assign the spectral intensity value at a point in the model while the value of 0 described the lack of taking into account this value in a substituent of the compound.

A pool of 100 chromosomes was tested using a random choice of the points on the surface. For each chromosome a CP-ANN model was obtained and simultaneously

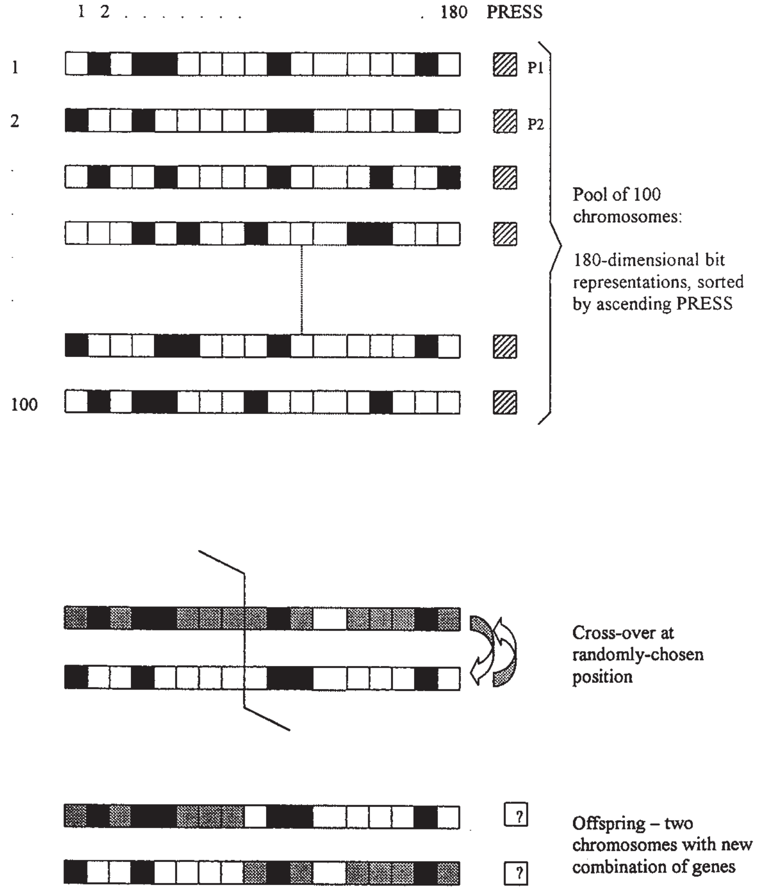
1 2 . . . . 180 PRESS
1
2
100
P1
P2
Pool of 100 chromosomes:
180-dimensional bit representations, sorted by ascending PRESS
Cross-over at randomly-chosen position
?
?
Offspring – two chromosomes with new combination of genes

Figure 4. Schematic representation of GA procedure. The 180-dimensional chromosomes are ordered according to the fitness function obtained from a CP-ANN model (PRESS of biological activity expressed as log MIC value of 10 test compounds). The cross-over event is shown on the first two chromosomes. Black and white squares indicate values 1 or 0 for individual choromosomes' compounds (genes). In order to present the difference between two parent chromosomes which produce two offspring chromosomes the white genes of one of them were labeled gray. See Ref. 12 and text for additional explanation.

the reduction of vector length was performed. These vectors with reduced length in the training set of compounds were used for training the CP-ANN. The quality of each such chromosome representation was again determined by PRESS in the test set. The best chromosomes were crossed over (mutated) and in the new pool of 100 chromosomes the new pattern representation was tested for quality. This procedure was repeated 200 times and thus in each generation a more representative pattern was obtained. Finally, the resulting optimized model CP-ANN incorporates the ability to predict the antibiotic activity of the compound.

In order to systematically vary the variables which determine the genetic algorithm (size of ANN, the representation length, number of iterations (epochs), speed of learning, number of generations in the algorithms, choice of bits out of 180 possible and mutation probability) SIMPLEX optimization was used. This method proved very useful for simultaneous optimization of a small number of variables.[7]

Five ANN and GA parameters were adjusted in the course of Simplex optimization: number of neurons in the ANN (range from 4×4 to 8×8), learning rate constant (from 0.1 to 0.9), number of epochs for training (from 160–640), number of starting bits in each chromosome turned on (from 3–30) and elitist rate (number of survivals from 3–30). Other parameters were constant: numbers and selection of objects in the training, test, and control sets for ANN modeling, number of 180-bit chromosomes in the pool (100), number of generations in each GA run (200), probabilities for mutation (0.005), cross-over was made exactly once at each 180 genes (bits) long chromosome.

For each Simplex point, defined by the above five parameters, a complete GA optimization consisting of 200 generation of the 100 chromosome pool each having 180 bits was made. The fitness function from which the survival chance of each chromosome was calculated was the RMS of the particular CP ANN model determined by the differences between the predictions and actual activities on the control set of 10 compounds. Each CP ANN model was based on different number of intensities in the spectrum-like representation. The selection of the intensities is determined by the bits turned »on« in the chromosome which in turn depends on the cross-over, mutation and selection procedure during the next chromosome pool generation of the GA.

In order to determine one optimization criterion of a single Simplex point 20,000 CP ANN models were generated and tested. For each GA run all CP ANNs have the same number of neurons, but different numbers of weight. Hence, the number of weights to be adjusted varied from 3×3×3 = 27 to 3×3×30 = 270 in the case of the smallest 3×3 ANN to 8×8×3 = 192 to 8×8×30 = 1920 in the case of the largest 8×8 ANN, respectively. To achieve the convergence for each ANN at least 160 epochs of training containing of 32 spectra were employed.

The GA and ANN computer routines were integrated into a single home made PC resident Fortran program. Due to its extreme simplicity the Simplex determination of the parameters for each next point were made by hand. Altogether 252 Simplex points or simplex movements from six different starting positions were tested. The contraction of simplex close to the local optima was made by a factor of 0.5.[45]

This means that in the entire optimization procedure about 5,000,000 CP ANN models were inspected.

RESULTS AND DISCUSSION

Model Optimization Using the ANN Approach

In Figure 5 the spectrum-like representation for is shown for molecule **1** of the data set. The unreduced length of this representation gives the complete chemical structure of the molecule as a 2D plot in which intensities of each component give the presence or absence of a substituent in given position on the tricylic skeleton.

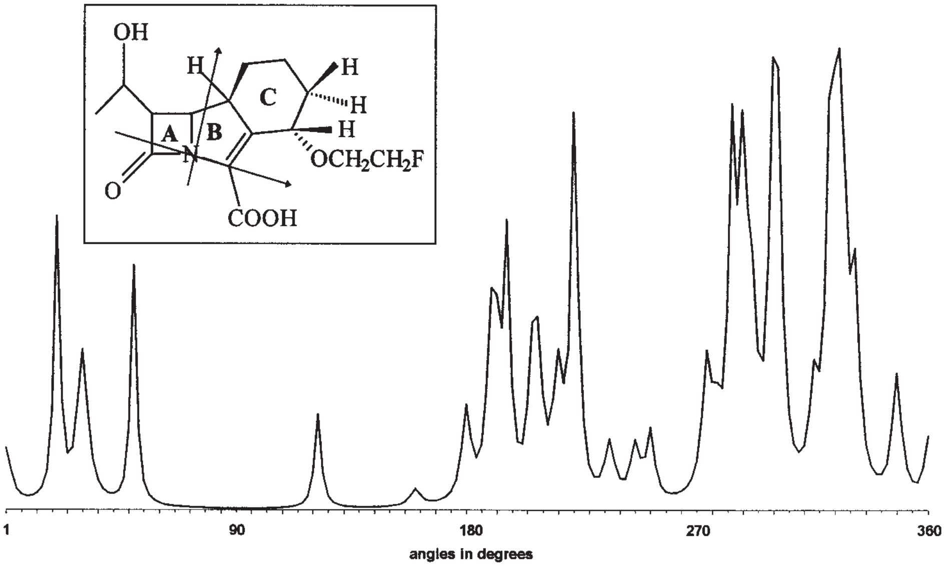

Figure 5. The spectrum-like representation for molecule **1** of the data set (see also Table I: $R_5 = OCH_2CH_2F$).

The selection of most relevant variables for the structure-activity relationship using a combination of ANN and GA algorithms as outlined above in the Methods section has two goals: (i) to reduce the dimension of the representation vectors (in our case, from 180 to 10–20) and (ii) to correlate the selected variables with a possible biochemical mechanism underlying antibiotic bioactivity of the tricylic carbapenem analogues in the series.

The accuracy of the results shows that this approach can result in a valuable description of the relationship between structure and bioactivity.

Figure 6 shows how selection of the training set on the basis of CV-LOO (leave one out) was performed on the experimental data set of 52 molecules and with 180 long representation of their structure. In order to select the compounds for the training set in an optimal way, a compound causing a large prediction error in LOO CV should be included in the training-set since it contains structural and property information that is more important in comparison with other compounds from the data set.

The final model is thus sensitive to the elimination of such unique compounds from the training set. The same division of the compounds into training and test sets was used to calculate the optimization criterion in the GA step which was then applied in order to select variables from the 180-dimensional structure representation vectors.

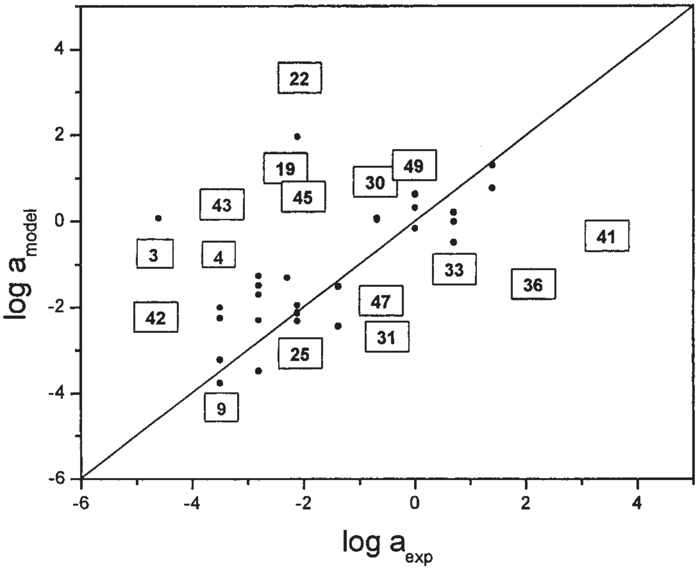

Figure 6. Selection of the training set on the basis of CV-LOO (leave one out) on the set of 52 molecules.

Thus, this intermediate CP ANN model generated from the non-reduced 180 dimensional structure representation provides the proof of concept that antibiotic activity for compounds in the test set can be predicted with the reasonable accuracy. It has to be stressed that these results are not describing the final CP ANN model. The ANN parameters are further changed during the GA procedure and these parameters then describe the final CP ANN model for the reduced representation.

Reducing the Representation Using GA

By application of the genetic algorithm procedure a selection of the variables (vector components) with the largest influence on the prediction ability is obtained, *i.e.* those having the largest influence on antibiotic activity of the compounds .

The optimization criterion in the GA algorithm was the correlation of the experimental MIC values for 10 test compounds with those predicted by a CP ANN model which was trained with 32 compounds. As shown in the Table II, 12 to 19 variables (denominated as »bits«, see also description of a genetic algorithm procedure in the Methods section) out of 180 were selected. The corresponding correlation coefficients are also shown in the Table II.

TABLE II

Summary of reduced representations given by GA procedure for the 20 experiments performed in the present work. The reduced representations are chosen with 12 to 19 bits from the full spectral length of 180 bits.

No	R_{train}	RMS_{train}	R_{test}	RMS_{test}	Bits	$\frac{RMS_{train}}{R_{train}} \times \frac{RMS_{test}}{R_{test}}$
1	0.990	0.220	0.876	0.678	15	0.17
2	0.995	0.196	0.775	0.802	12	0.20
3	0.968	0.394	0.923	0.520	16	0.23
4	0.996	0.162	0.648	0.979	13	0.25
5	0.989	0.227	0.788	0.972	12	0.28
6	0.901	0.909	0.943	0.437	16	0.47
7	0.981	0.575	0.818	0.911	16	0.65
8	0.981	0.575	0.818	0.911	17	0.65
9	0.981	0.577	0.818	0.910	16	0.65
10	0.956	0.534	0.800	1.012	14	0.71
11	0.997	0.205	0.926	3.628	16	0.81
12	0.980	0.563	0.787	1.187	14	0.87
13	0.998	0.159	0.807	5.183	17	1.02
14	0.993	0.257	0.939	4.298	19	1.19
15	0.400	1.176	0.926	0.507	14	1.61
16	0.484	1.149	0.915	0.690	13	1.79
17	0.643	1.009	0.970	1.541	15	2.49
18	0.511	1.183	0.988	1.714	14	4.01
19	0.524	1.559	0.402	1.784	17	13.20
20	0.516	1.253	0.212	1.331	15	15.17

The best correlation was obtained if the molecular structure was represented by the 15 variables. This reduced model gave the following statistical values for the training and test set, respectively:

$$R_{train} = 0.99,\ RMS_{train} = 0.22,\ R_{test} = 0.88,\ RMS_{test} = 0.68 \text{ and}$$

$$(RMS_{train} / R_{train}) \times (RMS_{test} / R_{test}) = 0.17 \ .$$

Using the resulting reduced representation the final correlation results are given in Figure 7. The correlation of experimental biological activity with computed bioactivity is shown for the training set ($N = 32$, symbol Δ,) and for the test set ($N = 11$, symbol ←). The final predictions for the non-biased set of 10 compounds (prediction set) which were excluded from any optimization since the start of the procedure are given in Figure 7 with symbols O.

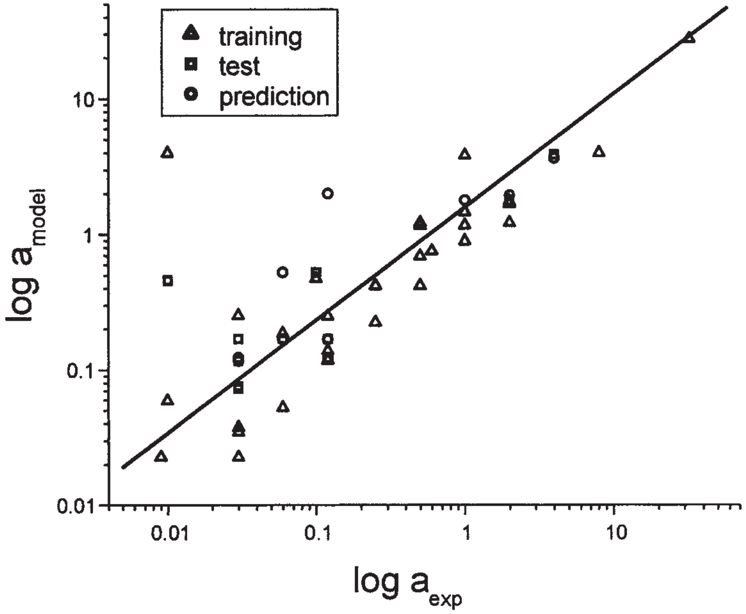

Figure 7. Correlation between computed and experimental bioactivity for the data set of 52 tricyclic carbapenems: predicted values are given by using the model based on the reduced representation obtained by the GA (see also Table II).

The reduced representation is expressed as a chromosome with the best fit value given by the PRESS value. In our case the chromosome with the best fit value of 0.17 has 15 genes turned to 1, which means that 15 of the selected variables contain a large amount of information for predicting the biological activity of the compounds under investigation. These 15 variables were analysed and found to originate from atoms present in substituents at position R_4 and R_5. Not all substituents determined by GA from the reduced representation were represented with equal frequency. Those that appeared the most frequently in the reduced representations (the data are given for the best 7 representations) are shown in Figure 8.

Thus, by conversion of the most populated angular frequencies in the spectrum-like structure representation yielded by this computational proce-

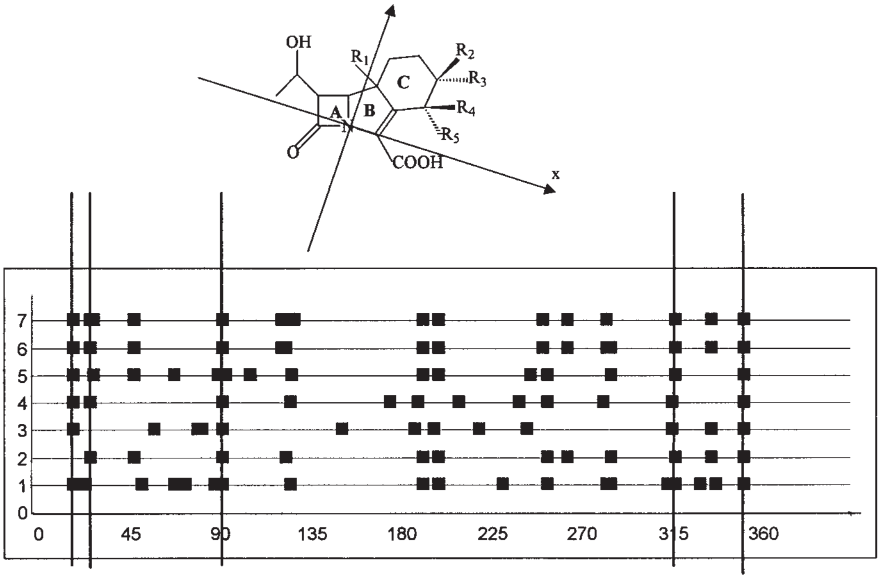

Figure 8. The most relevant directions with respect to the skeleton of the tricyclic carbapenem (coordinate system is shown in the insert). Seven best representations are shown. The important directions are in the 1st (0–90 degrees) and 4th (270–360 degrees) quadrant. See text for discussion.

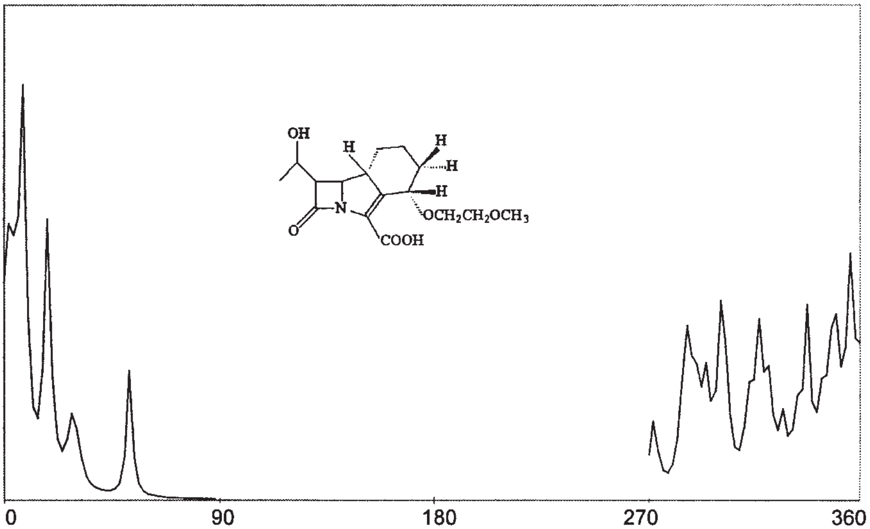

Figure 9. Molecule **41** with best bioactivity MIC > 32.0 μmol in the series as represented with the spectral-like and reduced representation obtained by the GA procedure.

dure into three dimensional molecular structure we obtain a valuable insight into structure-activity relationship of tricyclic carbapenems. Molecule **41** with best bioactivity MIC > 32.0 μmol in the series as represented with the spectral-like and reduced representation obtained by the GA procedure is shown in Figure 9. The substituent $OCH_2CH_2OCH_3$ in position R_5 covers spectral directions in the 1st and 4th quadrant and the position of this substituent in three-dimensional space is shown in Figure 8 above.

Furthermore, such procedure appears to provide a valuable model for design of novel tricyclic carbapenems based on correlating the experimental antibiotic bioactivities with their structure.[35,45]

CONCLUSIONS

We have used a variety of the artificial inteligence armory to obtain the model for correlation of experimental antibiotic bioactivity in a series of 52 tricyclic carbapenems with their threedimensional structure. A combination of ANN and GA based on spectrum-like representation of molecules in the data set yields good statistical values for this description. Such a model can be used for design of novel inhibitors of the target enzyme in question, D,D peptidase, whose active site is inhibited in order to stop the formation of the outer cell membrane in the bacterial cells.

Acknowledgements. – This work was supported in part by the Ministry of Science and Technology of Slovenia grant No. P0-0508-0104. Our thanks are due to Dr. Marjana Novič (National Institute of Chemistry) for many valuable discussions.

ABBREVIATIONS

ANN – artificial neural network
CP – counter-propagation
GA – genetic algorithm
PRESS – predicted residual errors sum of squares
MIC – minimal inhibitory concentration

REFERENCES

1. M. L. Cohen, *Science* **257** (1992) 1050–1055.
2. J. Davies, *Science* **264** (1994) 375–381.
3. C. M. Hansch, D. Hoekman, and H. Gao, *Chem. Rev.* **96** (1996) 1045–1075.
4. K. Gubernator and H. J. Boehm, *Rational Design of Bioactive Molecules*, in: K. Gubernator and H. J. Boehm (Eds.), *Structure-based Ligand Design,* Methods and Principles in Medicinal Chemistry, Vol. 6, Wiley-VCH, Weinheim, 1997, pp. 1–11.

5. M. Karelson, V. S. Lobanov, and A. R. Katritzky, *Chem. Rev.* **96** (1996) 1027–1043.
6. R. D. Cramer, S. A. DePriest, D. E. Patterson, and P. Hecht, *The Developing Practice of Comparative Molecular Field Analysis,* in: H. Kubinyi (Ed.), *3D QSAR in Drug Design Theory, Methods and Applications*, ESCOM Leiden, 1993, pp. 443–485.
7. J. Zupan and J. Gasteiger, *Neural Networks in Chemistry and Drug Design,* Wiley-VCH, Weinheim, 1999.
8. J. Devilers, *Strengths and Weaknesses of the Backpropagation Neural Network in QSAR and QSPR Studies*, in: J. Devillers (Ed.), *Genetic Algorithms and Molecular Modeling*, Academic Press, New York, 1996, pp.1–21.
9. J. Zupan, M. Novič, and I. Ruisánchez, *Chemom. Intell. Lab. System* **38** (1997) 1–23.
10. K. Illgen, T. Enderle, C. Broger, and L. Weber, *Chemistry & Biology* **7** (2000) 433–441.
11. M. Novič, Z. Nikolovska-Coleska, and T. Šolmajer, *J. Chem. Inf. Comp. Sci.* **37** (1997) 990–998.
12. G. Mlinšek, M. Novič, M. Hodoščček, and T. Šolmajer, *J. Chem. Inf. Comp. Sci.* **41** (2001) 1286–1294.
13. REGISTRY Manual, Chemical Abstracts Service, Columbus OH, USA, January 1998.
14. H. Schulz and U. Georgy, *From CA to CAS online; Databases in Chemistry,* Springer-Verlag, 1994.
15. D. Andreotti, T. Rossi, G. Gaviraghi, D. Donati, C. Marchioro, E. Di Modugno, and A. Perboni, *Bioorg. Med. Chem. Lett.* **6** (1996) 491–496.
16. S. Hanessian, M. J. Rozema, G. B. Reddy, and J. F. Braganza, *Bioorg. Med. Chem. Lett.* **5** (1995) 2535–2540.
17. D. Andreotti, T. Rossi, and C. Marchioro, *Bioorg. Med. Chem. Lett.* **6** (1996) 2589–2594.
18. S. Biondi, E. Piga, T. Rossi, and G. Vigelli, *Bioorg. Med. Chem. Lett.* **7** (1997) 2061–2066.
19. S. Gehanne, E. Piga, D. Andreotti, S. Biondi, and D. Pizzi, *Bioorg. Med. Chem. Lett.* **6** (1996) 2791–2794.
20. D. Andreotti, S. Biondi, R. Di Fabio, D. Donati, E. Piga, and T. Rossi, *Bioorg. Med. Chem. Lett.* **6** (1996) 2019–2024.
21. S. Hanessian, A. M. Griffin, and M. J. Rozema, *Bioorg. Med. Chem. Lett.* **7** (1997) 1857–1862.
22. R. Di Fabio, D. Andreotti, S. Biondi, G. Gaviraghi, and T. Rossi, *Bioorg. Med. Chem. Lett.* **6** (1996) 2025–2030.
23. A. Perboni, D. Donati, and G. Tarzia, *Eur. Pat. Appl*. **1992**, EP 0502 468 A1.
24. S. Biondi, D. Andreotti, T. Rossi, R. Carlesso, G. Tarzia, and A. Perboni, *Eur. Pat. Appl*. **1992**, EP 0502 464 A1.
25. A. Perboni and G. Sbampato, *PCT Int. Appl.* **1995**, WO 9523149 A1.
26. S. Biondi, G. Gaviraghi, and T. Rossi, *Bioorg. Med. Chem. Lett.* **6** (1996) 552–528.
27. M. E. Tranquillini, G. L. Araldi, D. Donati, G. Pentassuglia, A. Pezzoli, and A. Ursini, *Bioorg. Med. Chem. Lett.* **6** (1996) 1683–1688.
28. Y. Iso and Y. Nishitani, *Heterocycles* **48** (1998) 2287–2308.
29. R. Di Fabio, A. Feriani, G. Gaviraghi, and T. Rossi, *Bioorg. Med. Chem. Lett.* **5** (1995) 1235–1240.
30. M. Sendai and T. Miwa, *Eur. Pat. Appl.* **1990**, EP 0422 596 A2.

31. M. Sendai and T. Miwa, *Eur. Pat. Appl*. **1992**, EP 0507 313 A1.
32. F. P. Dininno, *PCT Int. Appl.* **1995**, WO 9503700 A1.
33. F. P. Dininno, *PCT Int. Appl*. **1995**, WO 9503699 A1.
34. A. Morimoto, N. Choh, and N. Noguci, *Eur. Pat. Appl*. **1987**, EP 0249 909 A1.
35. A. Copar, T. Šolmajer, B. Anzič, T. Kuzman, T. Mesar, and D. Kocjan, P*CT Int. Appl.* WO 98/27094.
36. G. S. Babini, M. Yuan, and D. M. Livermore, *Antimicrob. Agents. Chemother.* **42** (1998) 1168–1175.
37. E. Di Modugno, I. Erbetti, L. Ferrari, G. Galassi, S. M. Hammond, and L. Xerri, *Antimicrob. Agents Chemother.* **38** (1994) 2362–2368.
38. J. Zupan and M. Novič, *Anal. Chim. Acta* **384** (1997) 409–418.
39. Spartan User's guide, Version 4.0, Wavefunction, Inc., Irvine, CA, USA, 1995.
40. J. Zupan, M. Vračko, and M. Novič, *Acta Chim. Slov.* **47** (2000) 19–37. Note: the paper is available on the www at: http://www.acta.chem-soc.si/47/47–1–19.html
41. L. Eriksson and E. Johansson, *Chem. Intell. Lab. System* **34** (1996) 1–19.
42. J. Zupan and M. Novič, *Anal. Chim. Acta* **388** (1999) 243–250.
43. R. Hecht-Nielsen, *Appl. Optics* **26** (1987) 4979–4984.
44. T. Šolmajer and A. Copar, *Seventh* β*-lactamase Workshop*, Holy Island (UK), 5–9 April, 1998.
45. K. W. C Burton and G. Nickless, *Chemom. Intell. Lab. Syst.* **1** (1987) 135–149.

SAŽETAK

Kvantitativni odnos strukture i aktivnosti troprstenih karbapenema: Primjena metoda umjetne inteligencije za predviđanje bioaktivnosti

Mira Lebez, Tom Šolmajer i Jure Zupan

Rezistencija bakterija na antibiotike privukla je pažnju znanstvenika koji rade na razvoju novih terapeutskih spojeva. Najveća skupina antibiotika na tržištu su penicilini i cefalosporini koji imaju istu *β*-laktamsku strukturnu jedinicu. Njihovi nedavno otkriveni troprsteni analozi pokazali su izvanrednu bioaktivnost prema širokom spektru bakterijskih vrsta. U nizu od 52 troprstenih karbapenema prikazanim pomoću 180'dimenzionalne »spectru nalik« reprezentacije autori su studirali odnos strukture i aktivnosti s pomoću neuronske mreže. Prikaz molekulske strukture s pomoću vrijednosti spektralnih intenziteta poslužili su kao ulazni podatci za neuronsku mrežu (CP-ANN). SIMPLEX optimizacija provedena je da se dobije najbolji ANN model, a zatim je upotrijebljen genetički algoritam za simultano smanjivanje broja varijabli. Tako je provedeno traganje za supstituentom koji pretežno utječe na eksperimentalnu bioaktivnost. Dobiveni CP-ANN model predviđa bioaktivnosti s koeficijentom korelacije od 0,88, u rasponu vrijednosti od preko 4 reda veličine. Lista supstituenata koju su autori dobili njihovim automatskim postupkom usporediva je s podatcima dobivenima proteinskom kristalografijom *β*-laktamskih inhibitora u kompleksu s enzimom D,D-peptidazom.