QSAR & Combinatorial Science QCS

Antibacterial Activities of Carbapenem Derivatives and Quantitative Structure – Activity Relationship for Drug Design

Jian Liu[a], Lu Zhou[a]* and Zhili Zuo[b]

[a] College of Chemical Engineering, Sichuan University, Sichuan, Chengdu 610065, China, E-mail: zhouluscu@163.com
[b] Centre for Biomedical and Life Sciences, Singapore Polytechnic, 139651 Singapore, Singapore

Keywords: Carbapenem, Drug design, GA-ANN, QSAR

Received: July 17, 2008; Accepted: May 15, 2008

DOI: 10.1002/qsar.200710104

Abstract

β-Lactams, a class of antibiotics including penicillins, cephems, monobactams, and carbapenems, is widely used. In this work, the Quantitative Structure – Activity Relationship (QSAR) models have been built by Partial Least Squares (PLS), Artificial Neural Network (ANN), and Genetic Algorithm optimized ANN (GA-ANN) to study the antibacterial activities of carbapenem derivatives. Of the three methods, GA-ANN is potentially useful in predicting QSAR properties of chemical agents. Furthermore, some novel penem derivatives are designed using classical bioisosterism strategy, and the antibacterial activities are predicted by the built GA-ANN model without any chemical or biological experiments. In conclusion, the niapenems and thiapenems show good antibiotic activities.

1 Introduction

β-Lactams, a class of antibiotics that includes penicillins, cephems, monobactams, and carbapenems [1], is widely used. Among these β-lactams, carbapenems show a broad spectrum of pathogens, including Gram-positive and Gram-negative bacterium, and also show good antibacterial effects. Carbapenems thus may be developed to the major classes of β-lactams [2]. The major difference between carbapenems and penicillins is that there is an unsaturated five-membered ring in carbapenems and a sulfur atom at C-1 instead of a carbon atom. Thienamycin is discovered in 1976 [3], which is a carbapenem derivative exhibiting good antimicrobial activity. Subsequently, more and more carbapenem derivatives have been synthesized. However, there are still some problems. The antibacterial activity against Methicillin-Resistant *Staphylococcus aureus* (MRSA), is particularly relatively weak. Thus, further structure-based drug design study is needed to discover new antibiotics.

Quantitative Structure – Activity Relationship (QSAR) models, mathematical equations relating chemical structure to their biological activities, give information that is useful for drug design and medicinal chemistry [4]. This method attempts to relate structural descriptors of molecules with their physicochemical properties and biological activities. It is widely used for the prediction of physicochemical properties in chemical, environmental, and pharmaceutical areas [5]. The success of QSAR approach can be explained by the structural determination of chemical properties, and the possibility to estimate the properties of new chemical compounds without any experiments [6]. Recently, many QSAR models which can successfully predict antimicrobial activity are reported [7 – 10]. The main steps of this method include data collection, molecular descriptors calculation and selection, correlation model development, and finally model evaluation.

A challenging problem in QSAR studies is the selection of the suitable modeling method. Different techniques have been used for establishing QSAR models including Multiple Linear Regression (MLR) [11, 12], Partial Least Squares (PLS) [13], and Artificial Neural Network (ANN) [14]. ANN has grown in popularity due to its ease of use and success in solving problems where complex nonlinear relationships exist [15 – 18]. However, the ANN has some limitations, such as overfitting and local optimum. Genetic Algorithm (GA) is a general evolutionary algorithm that can be used for optimization [19]. The combining GA with ANN (GA-ANN) can solve the problems that exist in ANN.

In this article, the QSAR models have been built by PLS, ANN, and GA-ANN to study the antibacterial activities of carbapenem derivatives, and compare the results obtained from the three methods. Of the three methods, GA-ANN is potentially useful in predicting QSAR properties of chemical agents. Then the classical bioisosterism strategy is applied to design some new penem derivatives, and their antibacterial activities are predicted using the built GA-ANN model without any experiments.

2 Materials and Methods

2.1 Biological Activity Data

The biological activity data of 83 carbapenem derivatives are extracted from the work of Azami *et al.* [20 – 25]. All the derivatives are tested for antibacterial activities *in vitro* against methicillin-suspectible *S. aureus* (MRSA). The biological activity data, inhibitory concentration (MIC) values, are determined using twofold dilution methods [26, 27]. The MIC values are converted to the logarithmic scale (pMIC) and then used as dependent variables as in QSAR analysis. The structural features of the compounds and their biological data are listed in Table 1.

2.2 Molecular Descriptors Calculation

Molecular structures of compounds were constructed in SYBYL7.3 (Tripos Associates, Inc., St. Louis, MO, USA) sketch molecule package [28]. The initial structures were first minimized using molecular mechanics with the MMFF94 force field [29] until the RMS of potential energy is smaller than 0.001 kcal/mol. Nearly 680 molecular descriptors, such as 0D, 1D, and 2D, were calculated using SYBYL. The brief description of those descriptors is presented in Table 2.

2.3 Division of the Dataset

In order to obtain a reliable QSAR model, an available dataset must be divided into the training set and prediction set. In this study, the division of a dataset into the training set and prediction set was performed using clustering techniques. A cluster sampling algorithm will focus on densely occupied region of the space and hence avoid region outliers. After the clustering process, the structure closest to the center of a cluster is selected as the representative structure. The dataset is divided into training set and prediction set (10%) by a *K*-means clustering algorithm clustering on descriptors (X) and biological activity (Y) values together [30]. Clustering on X and Y data together, rather than just on X, can cluster compounds according to all of the given information. This may lead to different prediction sets for different groups of indices but is appropriate when searching for the best model to represent a dataset [31 – 34].

2.4 Variable Selection

To build robust and accurate models, the models must be trained by a set of feature descriptors instead of all generated descriptors. The method of reducing the descriptors space is to extract features by building linear and nonlinear combinations of a lower dimension of the input features, which is called feature extraction. And this method extracts the information of the original descriptors into new variables by simple algorithms such as PCA [35].

In addition, in order to determine the variables which are significantly correlated with activity, regression coefficients (B) [36, 37], and Variable Importance for the Projection (VIP) [37, 38] of molecular descriptors are used to find the descriptor variables which are the most relevant to explain pMIC. High values of the regression coefficients signify that the descriptors are important to the regression. The VIP represents the value of each predictor in fitting the model for both predictor and response. If a predictor has a relatively small coefficient, then it is a prime candidate for deletion. The predictors with coefficients smaller than 0.01 are certainly suspect, and that with coefficients smaller than 0.05 probably do not contribute much.

2.5 Partial Least Squares

PLS is used to generate the linear models and it is performed with the MATLAB 7.0.1 (MathWorks, Inc.). PLS introduced by Wold [13] is well suited for problems with multicollinear predictor and response variables. The predictive values of the models are evaluated by Leave-One-Out (LOO) [49] cross-validation. The cross-validated coefficient, Q^2, is calculated using the following equation

$$Q^2 = 1 - \frac{\sum (y_i - y)^2}{\sum (y - y_{\text{mean}})^2} \quad (1)$$

where y_i, y, and y_{mean} are predicted, actual, and mean values of the target property (pMIC), respectively. To maintain the optimum number of PLS components and minimize the tendency to overfit the data, the number of components corresponding to the lowest PRESS value is to derive the final PLS regression model. PRESS is defined as

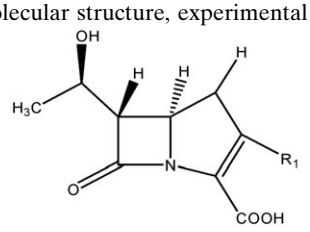
$$\text{PRESS} = \sum_{i=1}^{n} (y_i - y)^2 \quad (2)$$

In addition to the Q^2 and number of components, the squared correlation coefficients (R^2) of experimental pMICs *versus* predicted pMICs of training and prediction set, F-statistics, Standard Deviations (SD) and RMSE of training and prediction set are also computed.

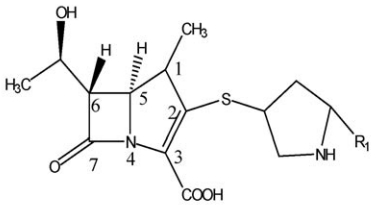
2.6 Artificial Neural Network

ANN is used to generate the nonlinear models and it is performed with the MATLAB 7.0.1 (MathWorks, Inc.). In this work, Back-Propagation ANN (BP-ANN) is used. In BP-ANN, "learning" is a supervised process that occurs with each cycle of "epoch" through a forward activation flow of inputs and the backward error propagation of weight adjustment [50]. In this work, gradient descent with

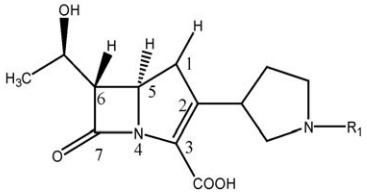
Table 1. Molecular structure, experimental and calculated antibiotic activities of carbapenems.



A



B



C

NO.	R_1	pMIC	Calculated pMIC (g/mol)		
			PLS	ANN	GA-ANN
A-1	2-Propenyl-pyrrolidine	3.486	3.810	3.656	3.735
A-2	2-Propenyl-pyrrolidine	3.466	3.761	3.793	3.668
A-3	2-(2-Me)-propenyl-pyrrolidine	3.787	3.854	3.694	3.735
A-4	2-(2-CH_2OCH_3)-propenyl-pyrrolidine	3.527	3.474	3.456	3.589
A-5	2-(2-CH_2SCH_3)-propenyl-pyrrolidine	3.547	3.702	3.663	3.728
A-6	2-(2-CH_2NH_2)-propenyl-pyrrolidine	3.510	3.836	3.658	3.589
B-7	CH_2OCH_2COOH	3.587	3.817	3.788	3.759
B-8[a]	$CH_2OCH_2NH_2$	3.903	4.126	4.146	4.075
B-9[a]	$CH_2OCH_2CONHCH_3$	4.122	4.188	4.118	4.031
B-10	$CH_2OCH_2CON(CH_3)_2$	3.932	4.329	4.175	4.144
B-11	$CH_2OC(CH_3)_2CONH_2$	4.233	4.139	4.043	4.038
B-12	$CH_2O(CH_2)_2OH$	3.888	3.926	4.138	4.071
B-13	$CH_2O(CH_2)_2NH_2$	4.489	4.116	4.337	4.266
B-14	$CH_2O(CH_2)_2NHCH_3$	4.505	4.229	4.226	4.342
B-15	$CH_2O(CH_2)_2N(CH_3)_2$	3.918	4.213	4.193	4.109
B-16	$CH_2O(CH_2)_2F$	4.191	3.972	4.027	4.115
B-17[a]	$CH_2OCH_2CHFCH_3$	3.605	4.127	4.015	3.708
B-18[a]	$CH_2O(CH_2)_2NHCONH_2$	4.222	4.183	4.277	4.293
B-19	3-$CH_2OCH_2CH_2$-1-Me-3H-imidazol-1-ium	4.558	4.624	4.624	4.608
B-20	2-$CH_2OCH_2CH_2$-1-Me-2H-pyrazol-1-ium	4.558	4.819	4.708	4.733
B-21	$CH_2O(CH_2)_4F$	4.234	4.246	4.275	4.165
B-22	$CH_2O(CH_2)_2Cl$	4.209	4.034	4.175	3.967
B-23	CH_2OCH_3	4.756	3.858	4.639	4.626
B-24	$C(NH)CH_3$	3.832	3.927	3.871	4.067
B-25	$CO(NH_2)_2$	4.186	4.174	4.051	4.120
B-26	6,7-dihydro-5H-pyrazolo[1,2,4]trizol-4-ylium	4.148	4.022	4.108	4.225
B-27	3-CH_2-1-Me-3H-imidazol-1-ium	4.212	4.245	4.169	4.154
B-28	1-CH_2-5-Me-1H-imidazo[1,2,4]pyrazol-5-ium	4.554	4.700	4.574	4.544
B-29	1-Me-3-ethyl-3H-imidazol-1-ium	4.528	4.772	4.777	4.780
B-30	1-Me-2-ethyl-2H-pyrazol-1-ium	4.528	4.784	4.724	4.657
B-31	4-Me-1-ethyl-1H-[1,2,4] triazol-4-ium	4.830	4.873	4.786	4.863
B-32	1-Ethyl-pyridinium	4.926	4.812	4.713	5.051
B-33	3-Ethyl-1-Me-3H-imidazol-1-ium	4.843	4.566	4.662	4.685
B-34	3-CH=$CHCH_2$-1-Me-3H-imidazol-1-ium	4.540	4.692	4.724	4.783
B-35	1-CH=$CHCH_2$-pyridinium	4.838	4.761	4.696	4.666
B-36	1-Me-3-ethyl-3H-imidazol-1-ium	4.528	4.594	4.680	4.762
B-37	1-CH_2CONH_2-3-ethyl-3H-imidazol-1-ium	4.557	4.877	4.664	4.704
B-38[a]	1-$(CH_2)_3NH_2$-3-ethyl-3H-imidazol-1-ium	4.871	4.941	4.828	4.880
B-39	2-CH_2OH-1-Me-3-ethyl-3H-imidazol-1-ium	4.558	4.348	4.338	4.548
B-40	5-CH_2OH-1-Me-3-ethyl-3H-imidazol-1-ium	4.841	4.678	4.638	4.544
B-41	4-CH_2OCH_3-Me-3-ethyl-3H-imidazol-1-ium	4.571	4.500	4.493	4.543
B-42	2-$CONH_2$-1-Me-3-ethyl-3H-imidazol-1-ium	4.872	4.651	4.630	4.720
B-43[a]	5-$CONH_2$-1-Me-3-ethyl-3H-imidazol-1-ium	4.872	4.820	4.769	4.753
B-44	5-CH_2CONH_2-l-3-Me-l-3-ethyl-3H-imidazol-1-ium	4.583	4.959	4.719	4.861
B-45	5-CH=$CHCONH_2$-1-Me-3-ethyl-3H-imidazol-1-ium	4.895	4.985	5.000	4.988
B-46	5-CN-1-Me-3-ethyl-3H-imidazol-1-ium	4.854	4.562	4.592	4.663
B-47	4-CH_2-1,3-dimethyl-3H-imidazol-1-ium	4.829	4.598	4.670	4.920
B-48	2-CH_2-1,3-dimethyl-3H-imidazol-1-ium	4.829	4.562	4.655	4.687
B-49	4-CH_2-1,2-dimethyl-2H-imidazol-1-ium	5.129	4.829	4.941	4.936

Table 1. (cont.)

NO.	R_1	pMIC	Calculated pMIC (g/mol)		
			PLS	ANN	GA-ANN
B-50	5-CH_2-1,2-dimethyl-2H-imidazol-1-ium	4.829	4.611	4.694	4.624
B-51	4-CH_2-1,3-dimethyl-3H-[1,2,3]triazol-1-ium	4.228	4.109	4.058	4.128
B-52	5-CH_2-1,4-dimethyl-1H-[1,2,4]triazol-4-ium	4.529	4.711	4.716	4.555
B-53	2-CH_2-1-Me-pyridinium	4.525	4.675	4.706	4.594
B-54	3-CH_2-1-Me-pyridinium	4.826	4.534	4.629	4.803
B-55	4-CH_2-1-Me-pyridinium	4.826	4.673	4.724	4.620
B-56[a]	4-CH_2-1-$(CH_2)_2$OH-3-Me-3H-imidazol-1-ium	4.558	4.641	4.687	4.450
B-57	4-CH_2-2-$(CH_2)_2$OH-1,3-dimethyl-3H-imidazol-1-ium	4.859	4.685	4.790	4.769
B-58	1-CH_2CONH_2-4-CH_2-3-Me-3H-imidazol-1-ium	4.570	4.781	4.522	4.557
B-59	5-CH_2-2-CH_2OH-1-Me-2H-pyrazol-1-ium	4.859	4.763	4.644	4.711
B-60	5-CH_2-4-$(CH_2)_2$OH-1,2-dimethyl-2H-pyrazol-1-ium	4.257	4.378	4.439	4.386
B-61	2-CH_2OH-4-CH_2-1-Me-2H-pyrazol-1-ium	4.544	4.873	4.616	4.744
B-62	3-CH_2OH-1,2-dimethyl-2H-pyrazol-1-ium	4.845	4.460	4.432	4.612
B-63	1-$(CH_2)_2$OH-pyridinium	4.541	4.607	4.545	4.667
B-64	2-CH_2OH-1-Me-pyridinium	4.842	4.440	4.469	4.665
B-65	1-CH_2CONH_2-pyridinium	4.855	4.933	4.701	4.752
B-66	2-$CONH_2$-1-Me-pyridinium	4.552	4.618	4.652	4.613
A-67	4-Piperidine	3.448	3.680	3.594	3.676
A-68	3-Piperidine	3.749	3.680	3.631	3.518
A-69	3-Pyrrolidine	3.425	3.719	3.734	3.363
A-70	3-Azetidine	3.402	3.699	3.793	3.538
A-71	S$(CH_2)_2$NHCH=NH	3.797	4.065	3.977	4.054
C-72	CH=NH	4.069	4.137	4.402	4.344
C-73	C(CH_3)=NH	4.692	4.218	4.526	4.445
C-74	CH=N(CH_3)	4.995	4.256	4.885	4.816
C-75	3,4-Dihydro-2H-pyrrole	4.426	4.513	4.497	4.729
C-76	C(CH_2OH)=NH	4.112	4.274	4.073	4.098
C-77[a]	C(CH_2OCONH_2)=NH	4.751	4.561	4.578	4.609
C-78	C(CH_2CONH_2)=NH	4.165	4.282	4.445	4.369
C-79[a]	C$(CH_2NHCONH_2)$=NH	4.731	4.768	4.676	4.510
B-80	5-CH_2-2-CH_2CONH_2-1-Me-2H-pyrazol-1-ium	4.871	4.738	4.731	4.639
B-81[a]	4-CH_2-2-CH_2CONH_2-1-Me-2H-imidazol-1-ium	4.871	5.051	4.729	4.760
B-82	4-CH_2-2-CH_2COOH-1-Me-2H-imidazol-1-ium	4.571	4.565	4.667	4.344
B-83	4-CH_2-$(CH_2)_2$OH-1-Me-3H-[1,2,3]triazol-1-ium	4.259	4.009	4.157	4.193

[a]The compounds using in the prediction set.

momentum is applied and the performance function is Root-Mean-Square-Error (RMSE), which is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\mathrm{PRESS}}{n}} \tag{3}$$

where n is the number of compounds.

For the basic gradient descent algorithm, the weights and biases are moved in the direction of the negative gradient of the performance function. Gradient descent with momentum often provides faster convergence [50]. Momentum can also help the network to overcome a shallow local minimum in the error surface and settle down at or near the global minimum [51].

2.7 Genetic Algorithm Optimized ANN Connection Weights

Genetic algorithm, which was first introduced in the early 1970s [19], becomes an important tool for optimizing functions. Genetic algorithm is a searching or optimizing algorithm based on Darwinian biological evolution principle.

The BP algorithm by which the network is trained begins with a random set of weights. GA uses survival of the fittest to learn connection weights in an ANN [52, 53]. In this study, we use a three layer (of nodes) ANN with 10 input, 6 hidden, and 1output nodes. For our architecture, the number of genes at each chromosome is defined as follows: $\zeta = 73[(10 \quad \text{inputs} + 1 \quad \text{threshold}) \times 6 \quad \text{hidden} + (6 \text{ hidden} + 1 \text{ threshold}) \times 1 \text{ output}]$. ζ is the number of genes in a population member, and ζ is a constant in all the population members.

Table 2. The calculated chemical descriptors used in this study.

Descriptor type	Molecular descriptor
Constitutional	Molecular weight, number of atoms, number of non-H atoms, number of heteroatoms, number of multiple bonds, number of aromatic bonds, number of functional groups, number of rings, number of H-bond donors, number of H-bond acceptors, *etc*.
Topological indices	Molecular connectivity: Chi indices [39], molecular shape: Kappa and Phi indices [40, 41], counts and complexity indices, topological state [42], shape, wiener [43], Shannon indices, vertex and edge counts, atom type electrotopological state indices [44], group type electrotopological state indices, randic connectivity indices [45], *etc*.
Chemical	Log *P*, hydration energy (E_{hydr}), Molar Refractivity (MR), Polarizability (Pol), molecular Surface Area (SA), molecular volume (V), *etc*.
Empirical descriptors	Dipole Moment (DM), HOMO and LUMO energies, heat of formation (H_{form}), total energy (E_{total}), electronic energy (E_{ele}), the local charges at each atom of the base unit of basic structure (LC_i), Most Positive Charge (MPC), Most Negative Charge (MNC), Sum of Squares of Charges (SSC), hardness (η), softness (S), electronegativity (χ), chemical potential (μ), electrophilicity (ω) [46], *etc*.
Charge topological indices	Galvez charge topological indices [47], mean topological charge indices order 1–10, global topological charge index, maximum, minimum, average and total charges, local dipole index, *etc*.
WHIM descriptors	Unweighted size [48], shape, symmetry and accessibility directional indices; size, shape, symmetry and accessibility directional indices weighted by atomic polarizability, atomic Sanderson electronegativity or atomic van der Waals volume; total size, shape symmetry and accessibility indices, *etc*.

During the connection weights optimization in GA, the chromosome and its fitness in the species represent a set of connection weights and predictive ability of the QSAR models, respectively. Each individual of the population is defined by a chromosome of binary values which represent a subset of connection weights. The population of the first generation is selected randomly. The operators used here are crossover and mutation (0–0.1% for mutation and 60–90% for crossover). The fitness score of each member of this new generation is evaluated again, and the reproductive cycle is continued until a desired number of generations or target fitness score is reached. Here, the fitness function is the RMSE.

3 Results and Discussion

As mentioned above, the aim of this work is to compare the results from PLS, ANN, and GA-ANN. In this section, the predictive abilities of the GA-ANN and two other models (PLS and ANN) are evaluated. All the data are scaled to unit variance [-1,1] before modeling. After the models have been built, the calculated values of pMIC need to be transferred back to the same units that are used for the original experimental values of pMIC for comparison purpose. To ensure a fair comparison, the same training set and prediction set is used for each model.

The 83 carbapenem derivatives are used in experiments. This dataset is divided into a training set of 73 compounds and a prediction set of 10 compounds by using a cluster technique [30–34]. The corresponding actual *versus* predicted values of the pMIC of all molecules studied by PLS, ANN, and GA-ANN are shown in Table 1.

All descriptors are included in the PLS model. LOO cross-validation is used to determine the number of PLS components. The PRESS statistic is based on the generated residuals. The cross-validation resulted in 10 LVs is the optimal number with a minimum PRESS and maximum Q^2. The result is obtained by the PLS model with R^2 (train) = 0.7039, RMSE (train) = 0.2509, R^2 (pred) = 0.7465, RMSE (pred) = 0.2265, Q^2 = 0.6373, and F = 17.1161. It indicates that the PLS model is improper in fitting this data. The plots of calculated *versus* observed values pMIC from the PLS are shown in Figure 1.

To predict antibacterial activities of carbapenem derivatives, descriptor selection is important for ANN model. The variables that produce the best model are selected using the values of regression coefficients and VIP as previ-

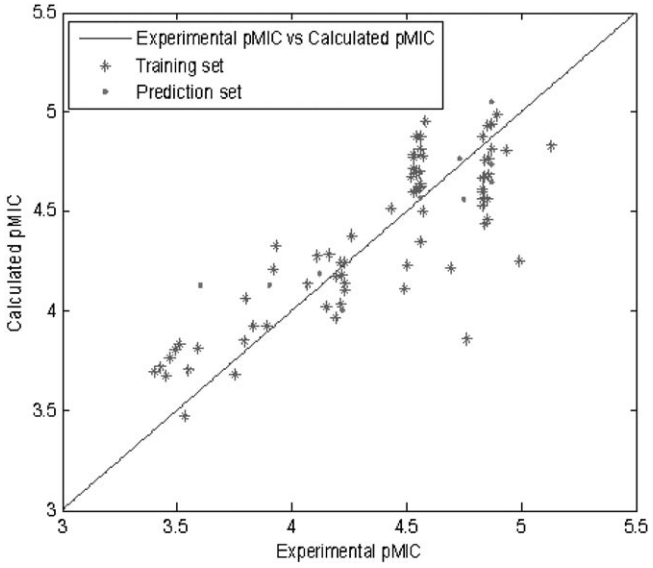


Figure 1. Correlation between experimental pMIC and PLS predicted pMIC for the training set and prediction set.

ously mentioned [36 – 38]. After the PLS analysis, it is noteworthy that these selected descriptors have no significant intercorrelation. Finally, ten descriptors are remained to build a nonlinear model. All of them have a relatively high regression coefficient (B), a high value of VIP and no significant intercorrelation. These descriptors are FPSA1 (partial positive surface area/total area), FNSA1 (partial negative surface area/total area), FNSA2 (total charge weighted partial negative surface area/total area), H-bond Acceptor (HA), electronic energy (E_{ele}), heat of molecular formation (H_{form}), n-octanol/water partition (log P), molecular dipole moment at z direction (DM_Z), softness (S), and chain of cycle terms of 5th order ($^5\chi_{CH}$).

The neural-network approach is especially suitable for analyzing complex nonlinear relationships between the activity and descriptor. The number of neurons in the hidden layer is an important factor determining the network's performance. Too many nodes may cause the network to memorize the dataset (overfitting); network with few nodes may be insufficient to use all the information from the dataset (underfitting). It is desirable to construct the network that generalize the patterns of the dataset rather than merely memorize them [54]. Previous studies conducted to determine the appropriate number of hidden units suggest that ρ, the ratio of number of data points to the number of adjustable weights in the neural network, may have a value between 1.8 and 2.3 [54, 55]. The range of ρ is used as a guideline for an acceptable number of neurons in the hidden layer when the increasing of hidden neurons does not improve the model anymore.

To solve the problems of overfitting and overtraining, is becoming an important factor for the improvement of generalization ability in neural network studies [56]. In the present study, a subdivision of the initial training set of 73 compounds into a learning set ($n=63$) and into a validation set ($n=10$) is done. The first set is used to train the network, whereas the second set is used to monitor the training process. The optimal training endpoint and network architecture are determined on the basis of this validation set. The network architecture and training endpoint which can make the model give the lowest RMSE and the maximal R^2, will be used for the predictions of the validation. In order to study the effect of network parameters on its performances, networks with different configurations are built. To avoid the results obtained by chance, the predictions are repeated 1000 times with different initial weights and the average pMIC values are calculated for each model. The network with six neurons in the hidden layer gives the best performance, as shown in Table 3. A sufficient training level is not reached with smaller number of neurons (<6) and overfitting exists with a larger number of neurons (>6) in the hidden layer. The optimal training ANN endpoint requires 8000 training epochs when the ANN architecture 10-6-1 is used. The architecture of 10-6-1 with minimum RMSE of training and prediction set is 0.1807 and 0.1943, respectively, and with maximum R^2 of the training and prediction set is 0.8556 and 0.8931, respectively. The architecture of 10-6-1 also gives results of $Q^2=0.7738$ and $F=42.6615$. So it is selected as the best ANN model. The plots of calculated *versus* observed values pMIC from the best nonlinear model of ANN are shown in Figure 2.

As a gradient search algorithm, ANN also has some limitations, such as overfitting, local optimum, and sensitivity to the initial values of weights. GA uses survival of the fittest strategy to learn connection weights in an ANN. The GA search process is terminated as the generation number reaches a predefined value of 200. The corresponding RMSE of the training set and the prediction set is 0.1647 and 0.1594, respectively, and R^2 is 0.8761 and 0.9565, respectively. Meanwhile the best prediction is obtained by the GA-ANN model with Q^2 of 0.8286 and F of 50.9114. In consequence, the BPNN prediction performance calculat-

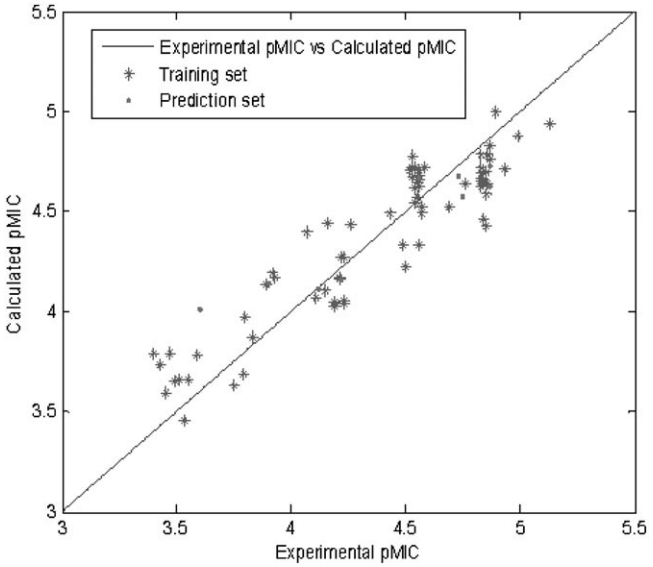


Figure 2. Correlation between experimental pMIC and ANN predicted pMIC for the training set and prediction set.

Table 3. The influence of different hidden neurons on the ANN performance.

ANN[a]	Training set		Prediction set	
	R^2	RMSE	R^2	RMSE
10-3-1	0.787	0.230	0.774	0.289
10-4-1	0.723	0.248	0.802	0.251
10-5-1	0.753	0.234	0.853	0.216
10-6-1	0.856	0.181	0.893	0.194
10-7-1	0.794	0.214	0.854	0.220
10-8-1	0.723	0.253	0.780	0.222
10-9-1	0.701	0.259	0.801	0.203
10-10-1	0.780	0.225	0.824	0.247
10-11-1	0.778	0.219	0.752	0.276
10-12-1	0.747	0.237	0.788	0.227

[a]Number of inputs-hidden-outputs neurons.

ed *versus* observed values pMIC from the GA-ANN is shown in Figure 3. The improvements of GA-ANN model over conventional PLS and ANN are shown in Table 4.

For classification, the RMSE (train), R^2 (train), Q^2, and F are listed in the following: GA-ANN gives the best results (0.1647, 0.8761, 0.8286, and 50.9114); ANN shows good results (0.1807, 0.8556, 0.7738, and 42.6615); and PLS shows poor results (0.2509, 0.7039, 0.6373, and 17.1161). For prediction, the RMSE and R^2 for the prediction set are listed in the following: GA-ANN gives the best results (0.1594 and 0.9565); ANN shows good results (0.1943 and 0.8931); and PLS provides a poor result (0.2265 and 0.7465).

The RMSE for the training set of linear model is larger than 0.25, the RMSEs (train) of the nonlinear models are all smaller than 0.2, the Q^2 of linear model is smaller than 0.7, and the F of linear model is smaller than 20, which indicates that the nonlinear model is better than the linear model during the training process. For the prediction ability, the RMSE (perd) of the best linear and nonlinear model is 0.1594 and 0.2265, respectively, which indicates that the predictive ability of the best nonlinear model is better than that of the best linear model. The reason is that the predictor variables of the carbapenem analogs data are nonlinearly correlated with response variables. Because the GA-ANN optimized ANN connection weights, the prediction performance is considerably improved and the result of GA-ANN is better than ANN. However, the neural-network approach is especially suitable for analyzing complex nonlinear relationships between the outputs and inputs. The result of ANN is better than PLS.

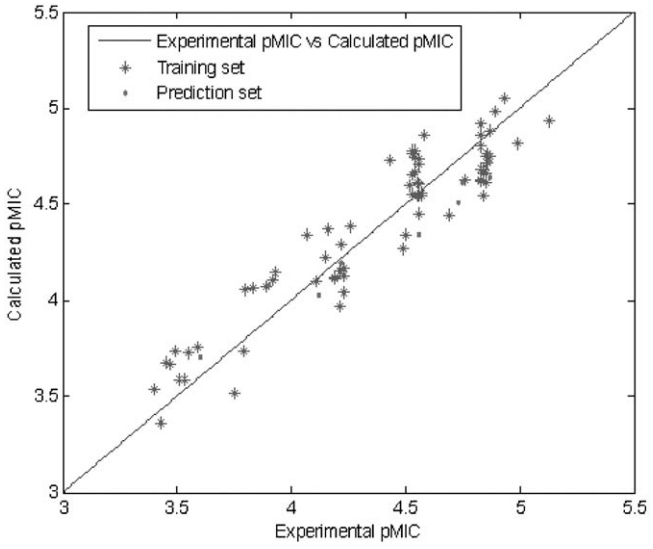


Figure 3. Correlation between experimental pMIC and GA-ANN predicted pMIC for the training set and prediction set.

Table 4. The results of GA-ANN, BPNN, and PLS.

Model	PLS	ANN	GA-ANN
R^2 (train)	0.704	0.856	0.876
Q^2	0.637	0.774	0.829
F	17.116	42.662	50.911
SD (train)	0.391	0.388	0.376
RMSE (train)	0.251	0.181	0.165
R^2 (pred)	0.747	0.893	0.957
RMSE (pred)	0.227	0.194	0.159
SD (pred)	0.358	0.309	0.296

4 Drug Design

β-Lactams antibiotics (penicillins and cephalosporins) are still widely used in the treatment of infectious diseases. The important pharmacophores (structural features responsible for molecule's biological activity) in these antibiotics are bicyclic molecules: cephams/cephems and penams/penems as shown in Figure 4. These pharmacophores contain β-lactams ring.

Penem derivatives substituted in various ways at any one or more of positions 3 and 6 were proposed previously, such as the carbon atom at position 1 has been replaced by a nitrogen atom (niapenems) [57], by an oxygen atom (oxapenems) [58], or by a sulfur atom (thiapenems) [59, 60]. To design novel penem derivatives with good antimicrobial activities, the classical bioisosterism strategy is applied without any chemical experiments [61]. The carbon atom at position 1 of the carbapenems is substituted by N, O, S, and the group of C-2 is unchanged which is used in carbapenem study (Figure 4). Therefore, we have 83 novel niapenem, oxapenem, and thiapenem derivatives, respectively.

The antimicrobial activities against MRSA of novel niapenem, oxapenem, and thiapenem derivatives can be predicted without any biological experiments. The steps are the following: first, the compounds were constructed and

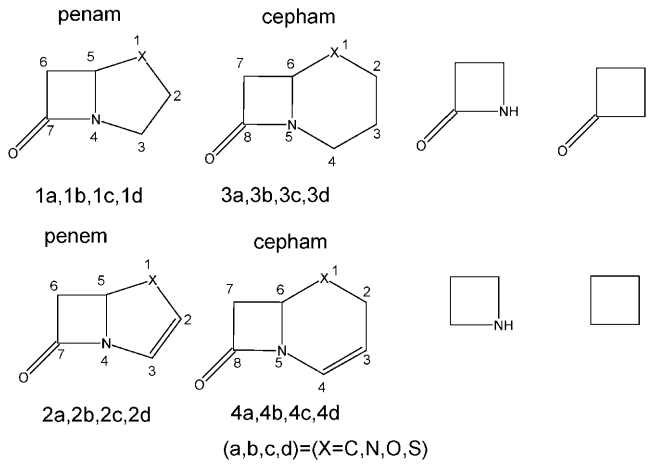


Figure 4. Substructures of cephams, cephems, penams, and penems.

Table 5. The calculated antibacterial activities of niapenems, thiapenems, and oxapenems.

No.	Experimental pMIC of carbapenems	Calculated pMIC		
		Niapenems	Thiapenems	Oxapenems
A-1	3.486	3.848	3.669	2.601
A-2	3.466	3.850	3.659	2.622
A-3	3.787	3.925	3.700	2.800
A-4	3.527	4.669	4.459	3.563
A-5	3.547	4.385	4.288	4.206
A-6	3.510	4.493	4.444	3.234
B-7	3.587	2.705	2.741	1.458
B-8	3.903	4.578	4.584	3.097
B-9	4.122	9.251	9.065	7.868
B-10	3.932	4.331	4.184	2.949
B-11	4.233	4.357	4.222	2.942
B-12	3.888	4.003	3.869	10.052
B-13	4.489	4.250	4.074	2.832
B-14	4.505	4.356	4.200	3.040
B-15	3.918	4.338	4.178	3.009
B-16	4.191	4.058	3.923	2.762
B-17	3.605	4.232	4.190	2.942
B-18	4.222	4.184	4.255	2.839
B-19	4.558	4.656	4.603	3.390
B-20	4.558	4.887	4.875	21.151
B-21	4.234	4.211	4.122	2.936
B-22	4.209	4.018	3.986	2.658
B-23	4.756	3.835	3.783	2.360
B-24	3.832	3.997	3.971	2.464
B-25	4.186	4.215	3.953	2.398
B-26	4.148	4.001	4.003	4.594
B-27	4.212	4.345	4.299	3.106
B-28	4.554	4.631	4.507	21.050
B-29	4.528	4.773	4.777	3.564
B-30	4.528	4.761	4.703	21.191
B-31	4.830	4.877	4.868	3.529
B-32	4.926	4.817	4.802	3.546
B-33	4.843	4.620	4.544	3.258
B-34	4.540	4.801	4.646	3.348
B-35	4.838	4.833	4.780	3.680
B-36	4.528	4.657	4.475	3.374
B-37	4.557	4.876	4.730	3.717
B-38	4.871	4.933	4.814	3.927
B-39	4.558	3.417	3.388	2.236
B-40	4.841	4.504	4.449	3.177
B-41	4.571	4.411	4.338	3.214
B-42	4.872	5.016	4.913	3.764
B-43	4.872	5.050	4.850	3.643
B-44	4.583	5.289	5.212	4.035
B-45	4.895	4.548	4.465	3.475
B-46	4.854	4.526	4.368	3.466
B-47	4.829	4.504	4.450	3.196
B-48	4.829	4.844	4.680	3.519
B-49	5.129	4.000	4.028	20.596
B-50	4.829	4.529	4.533	20.987
B-51	4.228	4.140	3.986	2.711
B-52	4.529	4.767	4.599	3.090
B-53	4.525	4.799	4.634	3.320
B-54	4.826	4.775	4.565	3.388
B-55	4.826	4.928	4.776	3.560
B-56	4.558	4.662	4.527	3.075
B-57	4.859	4.736	4.606	3.367
B-58	4.570	4.868	4.670	3.791
B-59	4.859	3.574	3.567	20.216

Table 5. (cont.)

No.	Experimental pMIC of carbapenems	Calculated pMIC		
		Niapenems	Thiapenems	Oxapenems
B-60	4.257	4.803	4.765	21.155
B-61	4.544	4.472	4.464	21.101
B-62	4.845	4.758	4.765	21.519
B-63	4.541	4.418	4.295	3.118
B-64	4.842	4.727	4.678	3.298
B-65	4.855	4.646	4.466	3.106
B-66	4.552	3.635	3.522	2.071
A-67	3.448	3.587	3.440	2.221
A-68	3.749	3.725	3.568	2.055
A-69	3.425	4.318	4.168	4.073
A-70	3.402	4.007	3.747	2.046
A-71	3.797	3.969	3.843	2.367
C-72	4.069	3.908	3.768	4.222
C-73	4.692	4.053	3.913	3.966
C-74	4.995	4.277	4.180	4.497
C-75	4.426	4.303	4.148	3.887
C-76	4.112	4.550	4.413	3.174
C-77	4.751	4.306	4.152	4.411
C-78	4.165	4.491	4.268	4.014
C-79	4.731	4.835	4.726	4.775
B-80	4.871	4.917	4.924	21.437
B-81	4.871	4.935	4.956	21.414
B-82	4.571	4.302	4.326	20.969
B-83	4.259	4.071	3.949	2.518

minimized using SYBYL. The initial structures were minimized using molecular mechanics with the MMFF94 force field until the RMS of potential energy is smaller than 0.001 kcal/mol. Second, 10 descriptors (FPSA1, FNSA1, FNSA2, HA, E_{ele}, H_{form}, log P, DM_Z, S,$^5\chi_{CH}$) were calculated and then used to predict the antimicrobial activities against MRSA of novel penem derivatives with the GA-ANN model constructed above. All of them had a relatively high regression coefficient (B), a high value of VIP, and no significant intercorrelation. Therefore, we can use the calculated descriptors to predict activities of novel penem derivatives without the need to synthesize and test them by experiments. Third, we utilize GA-ANN to predict the antimicrobial activities. Here, the input layer is changed by the calculated descriptors in the second step. And the output values are the antimicrobial activities of novel compounds. The calculated values are shown in Table 5.

As can be seen from Table 5, the antibiotic activities of niapenems and thiapenems are better than oxapenems. The reason for this is that the antibiotic activities of niapenems and thiapenems are similar to the carbapenems. As for oxapenem derivatives, there are only 4th, 72nd, 77th, 78th, and 79th compounds that show similar antibiotic activities to the carbapenems.

A major goal in pharmaceutical research is to design molecules that interact with specific biochemical pathways in living systems. A corresponding area in drug design aims at developing small organic molecules with a high affinity of binding against a given receptor. When a proper superposition of a set of ligands is available, the relevant chemical features of the ligands can be readily extracted in order to derive a pharmacophore model. In turn, the chemical features can be used to search for possible inhibitors in a ligand database. There are several superposition programs such as DISCO [62], GASP [63], and CATALYST [64]. In this research, the five penem compounds (the 7th, 9th, 20th, 60th, and 80th compounds) are selected for molecular superposition in SYBYL. The molecular superposition can be seen from Figures 5 and 6.

To distinguish the penem derivatives in Figure 5, the white atom at position 1 represents carbapenems, meanwhile the blue, red, and yellow atoms at position 1 represent niapenems, oxapenems, and thiapenems, respectively (Figure 4). The carbapenem is selected as a template molecule, and the others are superposed on it. In Figure 5, the effect of superposition of a set of carbapenems, niapenems, oxapenems, and thiapenems is poor. Compared to the antimicrobial activities of carbapenems, the other three penems derivatives show high deviation (as shown in Table 5). In Figure 6, the niapenems and thiapenems show good superpose on carbapenems. As a result, the antimicrobial activities are similar to the carbapenems, and the deviation of pMIC is only 0.3. In contrast to the niapenems and thiapenems, the effect of superposition of carbapenem and oxapenem derivatives is poor. Therefore, the deviation of antimicrobial activities is 17.00 (as shown in Ta-

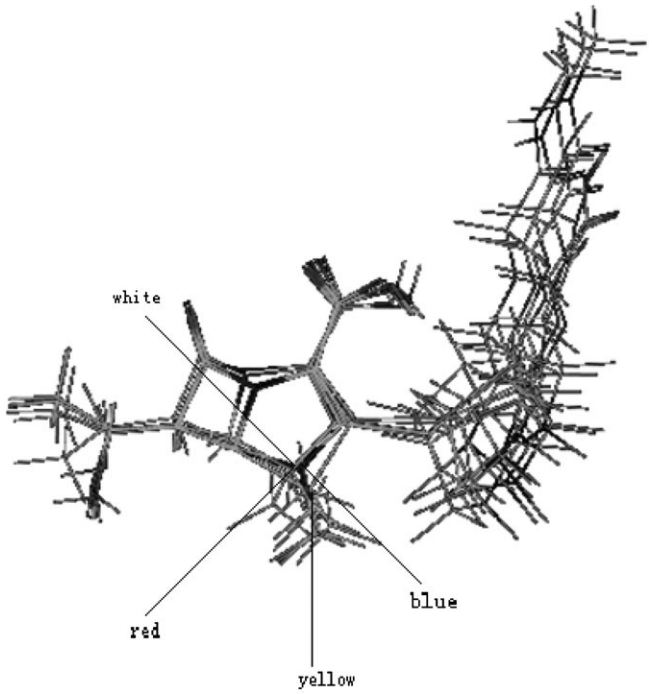


Figure 5. Superposing of the 7th and 9th penem compounds.

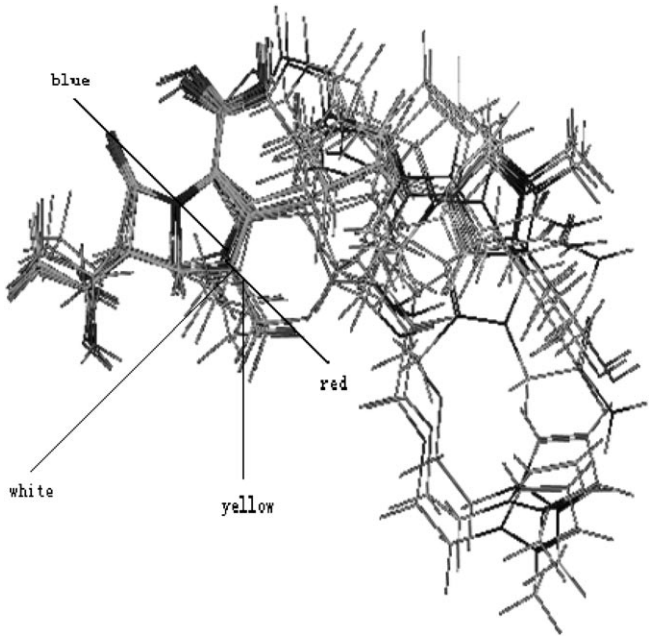


Figure 6. Superposing of 20th, 61st, and 80th penem compounds.

ble 5). We can conclude that QSAR studying and superposing a set of ligands may be an effective tool for drug design.

5 Conclusion

In this paper, the QSAR models PLS, ANN, and GA-ANN are employed to study the antibacterial activities of carbapenem derivatives. GA-ANN is potentially useful in predicting QSAR properties of chemical agents. The GA-ANN approach effectively optimizes ANN connection weights, and ANN which captures nonlinear relationships among predictor variables as well as with response variables through high-dimension feature mapping. Compared with the PLS, ANN has a better prediction performance. The linear regression model is found to be statistically valid.

To design novel penem derivatives with good antimicrobial activities, we apply classical bioisosterism strategy to design new structures and predict their antibacterial activities using the built QSAR model without any chemical or biological experiments. The antibacterial activities are related to compounds conformation. The niapenems and thiapenems show good superpose on carbapenems, and the antimicrobial activities are similar to the carbapenems. So QSAR studying and superposing a set of ligands may be an effective tool for drug design.

References

[1] R. B. Silverman, *The Organic Chemistry of Drug Design and Drug Action*, 2nd Edn., Elsevier, Amsterdam **2004**, pp. 1 – 614.
[2] G. P. Ellis, D. K. Luscombe (Eds.), *Progress in Medicinal Chemistry*, ElsevierScience, B. V., Amsterdam **1996**, pp. 99 – 145.
[3] J. S. Kahan, F. M. Kahan, R. Goegelman, S. A. Currie, M. Jackson, E. O. Stapley, T. W. Miller, A. K. Miller, D. Hendlin, S. Hernandez, H. B Woodruff, J. Birnbaum, *J. Antibiot.* **1979**, *32*, 1 – 12.
[4] D. Hadjipavlou-Litina, *Med. Res. Rev.* **1998**, *18*, 91 – 119.
[5] A. R. Katritzky, U. Maran, V. S. Lobanov, M. Karelson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1 – 18.
[6] O. Ivanciuc, T. Ivanciuc, A. T. Balaban, *Internet Electron. J. Mol. Des.* **2002**, *1*, 252 – 268.
[7] M. C. Bagchi, D. Mills, S. C. Basak, *J. Mol. Model.* **2007**, *13*, 111 – 120.
[8] F. J. Prado-Prado, H. Gonzalez-Diaz, L. Santana, E. Uriarte, *Bioorg. Med. Chem.* **2007**, *15*, 897 – 902.
[9] S. Alper-Hayta, E. Aki-Sener, B. Tekiner-Gulbas, I. Yildiz, O. Temiz-Arpaci, I. Yalcin, N. E. Altanlar, *Eur. J. Med. Chem.* **2006**, *41*, 1398 – 1404.
[10] H. Gonzalez-Diaz, E. Tenorio, N. Castanedo, L. Santana, E. Uriarte, *Bioorg. Med. Chem.* **2005**, *13*, 1523 – 1530.
[11] M. Fernandez, J. Caballero, *Bioorg. Med. Chem.* **2007**, *15*, 6298 – 6310.
[12] L. Saiz-Urra, M. P. Gonzalez, Y. Fall, G. Gomez, *Eur. J. Med. Chem.* **2007**, *42*, 64 – 70.
[13] S. Wold, in: H. van de Waterbeemd (Ed.), *Chemometric Methods in Molecular Design*, Wiley-VCH, Weinheim **1995**, pp. 195 – 218.
[14] G. Schneider, P. Wrede, *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175 – 222.

[15] M. Shamsipur, B. Hemmateenejad, M. Akhond, *Anal. Chim. Acta* **2002**, *461*, 147 – 153.
[16] J. Hunger, G. Huttner, *J. Comput. Chem*. **1999**, *20*, 455 – 471.
[17] S. Ahmad, M. M. Gromiha, *J. Comput. Chem*. **2003**, *24*, 1313 – 1320.
[18] C. L. Waller, M. P. Bradley, *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 345 – 355.
[19] D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, Reading, MA **1989**, pp. 40 – 64.
[20] A. Yamada, K. Hattori, S. Kuroda, T. Chiba, T. Kamimura, K. Sakane, *Bioorg. Med. Chem. Lett*. **1996**, *6*, 2705 – 2708.
[21] H. Azami, H. Tsutsumi, K. Matsuda, D. Barrett, K. Hattori, T. Nakajima, S. Kuroda, T. Kamimura, M. Murata, *Bioorg. Med. Chem*. **1997**, *5*, 2069 – 2087.
[22] H. Azami, D. Barrett, K. Matsuda, H. Tsutsumi, K. Washizuka, M. Sakusi, S. Kuroda, F. Shirai, T. Chiba, T. Kamimura, M. Murata, *Bioorg. Med. Chem*. **1999**, *7*, 1665 – 1682.
[23] H. Azami, D. Barrett, A. Tanaka, H. Sasaki, K. Matsuda, M. Sakurai, T. Terasawa, F. Shirai, T. Chiba, T. Matsumoto, S. Tawara, *Bioorg. Med. Chem*. **2001**, *9*, 961 – 982.
[24] K. Hattori, A. Yamada, S. Kuroda, T. Chiba, M. Murata, K. Sakane, *Bioorg. Med. Chem. Lett*. **2002**, *12*, 383 – 386.
[25] H. Azami, D. Barrett, A. Tanaka, H. Sasaki, K. Matsuda, M. Sakurata, T. Matsumoto, S. Iacara, T. Chiba, S. Sakane, *Bioorg. Med. Chem. Lett*. **1997**, *7*, 1409 – 1414.
[26] C. Betriu, S. Salso, A. Sanchez, E. Culebras, N. Gómez, I. Rodriguez-Arpach, J. Pteazo, *Int. J. Antimicrob. Agents* **2006**, *28*, 1 – 5.
[27] L. Jeyaseeli, A. Das Gupta, K. A. Kumar, K. Mazumdar, N. K. Dutta, S. G. Dosthan, *Int. J. Antimicrob. Agents* **2006**, *27*, 58 – 62.
[28] SYBYL 7.3 User Guide, Tripos, St. Louis, MO, USA **1999**.
[29] T. A. Halgren, *J. Comput. Chem*. **1996**, *17*, 490 – 519.
[30] F. R. Burden, D. A. Winkler, *J. Med. Chem*. **1999**, *42*, 3183 – 3187.
[31] J. T. Leonard, K. Roy, *QSAR Comb. Sci*. **2006**, *25*, 235 – 251.
[32] Y. Marrero-Ponce, M. Iyarreta-Veitia, A. Montero-Torres, C. Romero-Zaldivar, C. A. Brandt, P. Z. Ávila, K. Kirchgatter, Y. Machado, *J. Chem. Inf. Model*. **2005**, *45*, 1082 – 1100.
[33] M. P. Gonzalez, A. M. Helguera, M. A. Cabrera, *Bioorg. Med. Chem*. **2005**, *13*, 1775 – 1781.
[34] E. Molina, H. Gonzales-Diaz, M. P. Gonzalez, E. Rodriguez, Z. Uriarte, *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 515 – 521.
[35] H. Gohlke, F. Dullweber, W. Kamm, J. Marz, T. Kissel, G. Klebe, in: H. D. Hfltje, M. Sippl (Eds.), *Rational Approaches to Drug Design*, 13th European Symposium on QSAR, Barcelona, pp. 261 – 270.
[36] S. Wold, M. Sjostrom, L. Eriksson, *Chemometr. Intell. Lab. Syst*. **2001**, *58*, 109 – 130.
[37] G. Chong, C. H. Jun, *Chemometr. Intell. Lab. Syst*. **2005**, *78*, 103 – 112.
[38] S. Wold, *Chemometric Methods in Molecular Design*, Wiley-VCH, Weinheim **1995**, pp. 195 – 218.
[39] L. B. Kier, L. H. Hall, *J. Pharm. Sci*. **1981**, *70*, 583 – 589.
[40] L. B. Kier, *Quant. Struct. Act. Relat*. **1985**, *5*, 7 – 12.
[41] L. B. Kier, *Quant. Struct. Act. Relat*. **1985**, *4*, 109 – 116.
[42] L. H. Hall, L. B. Kier, *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, 1999.
[43] H. Wiener, *J. Am. Chem. Soc*. **1947**, *69*, 17 – 20.
[44] L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 1039 – 1045.
[45] M. Randic, *J. Am. Chem. Soc*. **1975**, *97*, 6609 – 6615.
[46] P. Thanikaivelan, V. Subramanian, J. Raghava Rao, B. Unni Nair, *Chem. Phys. Lett*. **2000**, *323*, 59 – 70.
[47] J. Galvez, R. Garcia, M. T. Salabert, R. Soler, *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 520 – 525.
[48] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, *J. Comput. Aided. Mol. Des*. **1997**, *11*, 79 – 92.
[49] S. Wold, *Technometrics* **1978**, *4*, 397 – 405.
[50] H. Demuth, M. Beale, *Neural Network Toolbox User's Guide*, The MathWorks, Natick, MA **1998**, pp. 137 – 194.
[51] M. T. Hagan, H. B. Demuth, M. H. Beale, *Neural Network Design*, PWS Publishing, Boston, MA **1996**, pp. 45 – 67.
[52] M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA **1996**, pp. 205 – 218.
[53] Z. Michalewicz, *Genetic Algorithms + Data Structures. Evolution Programs*, 3rd Edn, Springer-Verlag, Berlin, Heidelberg, NY **1996**, pp. 178 – 197.
[54] S. S. So, W. G. Richards, *J. Med. Chem.* **1992**, *35*, 3201 – 3207.
[55] T. A. Andrea, H. Kalayeh, *J. Med. Chem*., **1991**, *34*, 2824 – 2836.
[56] I. V. Tetko, D. J. Livingstone, A. I. Luik, *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 826 – 833.
[57] B. C. Ross, G. Johnson, *Eur. Pat. Appl.* **1981**, pp. 77. EP:31509.
[58] C. E. Jamieson, P. A. Lambert, I. N. Simpson, *Antimicrob. Agents Chem.* **2003**, *47*, 2615 – 2618.
[59] C. K. Murphy, D. M. Rothstein, *PCT Int. Appl.* **2007**, pp. 98. WO: 2007070084.
[60] J. O. Phillips, *PCT Int. Appl.* **2007**, pp. 44. WO: 2007070164.
[61] L. M. Lima, E. J. Barreiro, *Curr. Med. Chem*. **2005**, *12*, 23 – 49.
[62] Y. C. Martin, M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico, P. A. Pavlik, *J. Comput. Aided Mol. Design* **1993**, *7*, 83 – 102.
[63] Y. Kato, A. Inoue, M. Yamada, N. Tomioka, A. Itai, *J. Comput. Aided Mol. Design*. **1992**, *6*, 475 – 486.
[64] R. Bureau, C. Daveu, J. C. Lancelot, S. Rault, *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 429 – 436.

