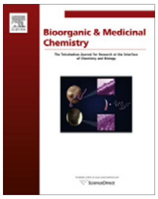
Bioorganic & Medicinal Chemistry 19 (2011) 2199–2210

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc



4D-QSAR analysis and pharmacophore modeling: Electron conformational-genetic algorithm approach for penicillins

Ersin Yanmaz[a], Emin Sarıpınar[b,*], Kader Şahin[c], Nazmiye Geçen[d], Fatih Çopur[b]

[a] Balıkesir University, Altınoluk Vacational College, Department of Chemistry, Balıkesir, Turkey
[b] Erciyes University, Science Faculty, Department of Chemistry, Kayseri, Turkey
[c] Bitlis Eren University, Faculty of Art and Sciences, Department of Chemistry, Bitlis, Turkey
[d] Siirt University, Faculty of Art and Sciences, Department of Chemistry, Siirt, Turkey

ARTICLE INFO

Article history:
Received 22 October 2010
Revised 1 February 2011
Accepted 19 February 2011
Available online 26 February 2011

Keywords:
Penicillins
4D-QSAR
Drug design
Pharmacophore
Electron conformational
Genetic algorithm

ABSTRACT

4D-QSAR studies were performed on a series of 87 penicillin analogues using the electron conformational–genetic algorithm (EC–GA) method. In this EC-based method, each conformation of the molecular system is described by a matrix (ECMC) with both electron structural parameters and interatomic distances as matrix elements. Multiple comparisons of these matrices within given tolerances for high active and low active penicillin compounds allow one to separate a smaller number of matrix elements (ECSA) which represent the pharmacophore groups. The effect of conformations was investigated building model 1 and 2 based on ensemble of conformers and single conformer, respectively. GA was used to select the most important descriptors and to predict the theoretical activity of the training (74 compounds) and test (13 compounds, commercial penicillins) sets. The model 1 for training and test sets obtained by optimum 12 parameters gave more satisfactory results ($R^2_{\text{training}} = 0.861$, $\text{SE}_{\text{training}} = 0.044$, $R^2_{\text{test}} = 0.892$, $\text{SE}_{\text{test}} = 0.099$, $q^2 = 0.702$, $q^2_{\text{ext1}} = 0.777$ and $q^2_{\text{ext2}} = 0.733$) than model 2 ($R^2_{\text{training}} = 0.774$, $\text{SE}_{\text{training}} = 0.056$, $R^2_{\text{test}} = 0.840$, $\text{SE}_{\text{test}} = 0.121$, $q^2 = 0.514$, $q^2_{\text{ext1}} = 0.641$ and $q^2_{\text{ext2}} = 0.570$). To estimate the individual influence of each of the molecular descriptors on biological activity, the *E* statistics technique was applied to the derived EC–GA model.

1. Introduction

β-Lactam antibiotics, known as penicillin-class antibiotics, are the most varied and widely used of all the different groups of antimicrobials, and about 100 different β-lactam antibiotics are used clinically in the antibacterial treatment of humans and animals.[1–3] Penicillins produce their bactericidal effects by inhibiting the synthesis of the peptidoglycan layer of bacterial cell walls. The basic structure of penicillin consists of a thiazolidine ring connected to a β-lactam ring, (the penicillin nucleus) and to which is attached a side chain (R). The penicillin nucleus is the chief structural requirement for biological activity and requires the presence of an acid residue on the thiazolidine ring for binding to the penicillin binding proteins.[4]

A QSAR study on β-lactam antibiotic derivatives has been reported using principal component and hierarchical cluster (PCA–HCA) analysis for 16 β-lactams.[5] In addition, the binding of 87 penicillins to human serum proteins was modeled with topological descriptors of molecular structure by Hall et al. using MDL-QSAR software.[6] The models indicated a combination of general and specific structure features that are important for estimating protein binding in this class of antibiotics. The predictive ability of the QSAR model was assessed using a test set of 13 commercial penicillin compounds.[6]

Rational drug design tries to establish a mathematical connection between the biological activity of a compound and some key molecular properties. The actual mathematical connection relies on statistics and relates biological activity to so-called molecular descriptors. The structural properties of the molecules are usually represented by a set of variables (descriptors), with the assumption that the molecule's activity is in some way related to the values of these variables.[7,8] All the descriptors generated for a specific molecule are not significant in modeling. The use of all available descriptors in model development causes dimensionality problems and diminishes the performance of a QSAR model, especially when non-linear algorithms are used in model development. Different methods for reduction are available in the literature. The genetic algorithms (GAs) have recently received much attention because of their ability to solve difficult problems in optimization.[9]

3D-QSAR methods are based on the detailed description of the local properties of each molecular structure. They are affected by the particular conformation adopted by a molecule, as well as to its orientation with respect to the other molecules. In 3D-QSAR studies, molecular alignment and 'active' conformation determination are so important that they affect the success of a model. One of

* Corresponding author. Tel.: +90 352 4374901, fax: +90 352 437493.
E-mail address: emin@erciyes.edu.tr (E. Sarıpınar).

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.02.035

the key steps in 3D-QSAR methodology is the selection of the active conformer for each compound in the series. For flexible molecules, this problem is the most difficult one and construction of the method with appropriate chemometric tools has been required.[10] Usually, a bioactive conformation of the ligand can be obtained from a structural determination of the ligand receptor complex by X-ray crystallography.[11]

The 4D-QSAR paradigm, which was proposed by Hopfinger et al. is a molecular modeling method that has proved both useful and reliable in the construction of quantitative 3D pharmacophore models for a set of ligand analogues when the geometry of the corresponding receptor is not known.[12] Basically, 4D-QSAR examines the conformational space of the molecular objects. These models are similar to 3D models, but as compared to them, the structural information is considered for a set of conformers (conditionally, the fourth dimension), instead of one fixed conformation.[13,14]

The electron conformational (EC) method presented by Bersuker et al. as one of the QSAR methods is aimed at searching rules for different activities prediction, based on the pharmacophores found previously by specific EC calculations.[15] For this purpose, a non-linear mathematical model which defines the relationship between bioactivity and the parameters was presented for bioactivity prediction using one conformer for compounds. This EC method has been recently applied to a variety of problems such as screening rice blast activity inhibitors, angiotensin converting enzyme inhibitors, group I metabotropic glutamate receptor agonists, inhibitors of human breast carcinoma, guanidino- and aminoguanidinopropionic acid analogues. A detailed description of the EC method has been adequately described in Bersuker's studies. Therefore, only the points relevant to this work are described here.[16–22]

Genetic algorithm (GA) is a heuristic search method used for identifying optimal solutions to a problem where the possible solution space is too large to be exhaustively enumerated.[23] GA has been widely used for feature optimization in QSAR models. This approach is able to elucidate structure–activity relationships by taking into account non-linear character of these relationships. The purpose of variable selection is to select the variables significantly contributing to prediction and to discard other variables by a fitness function.[24–26]

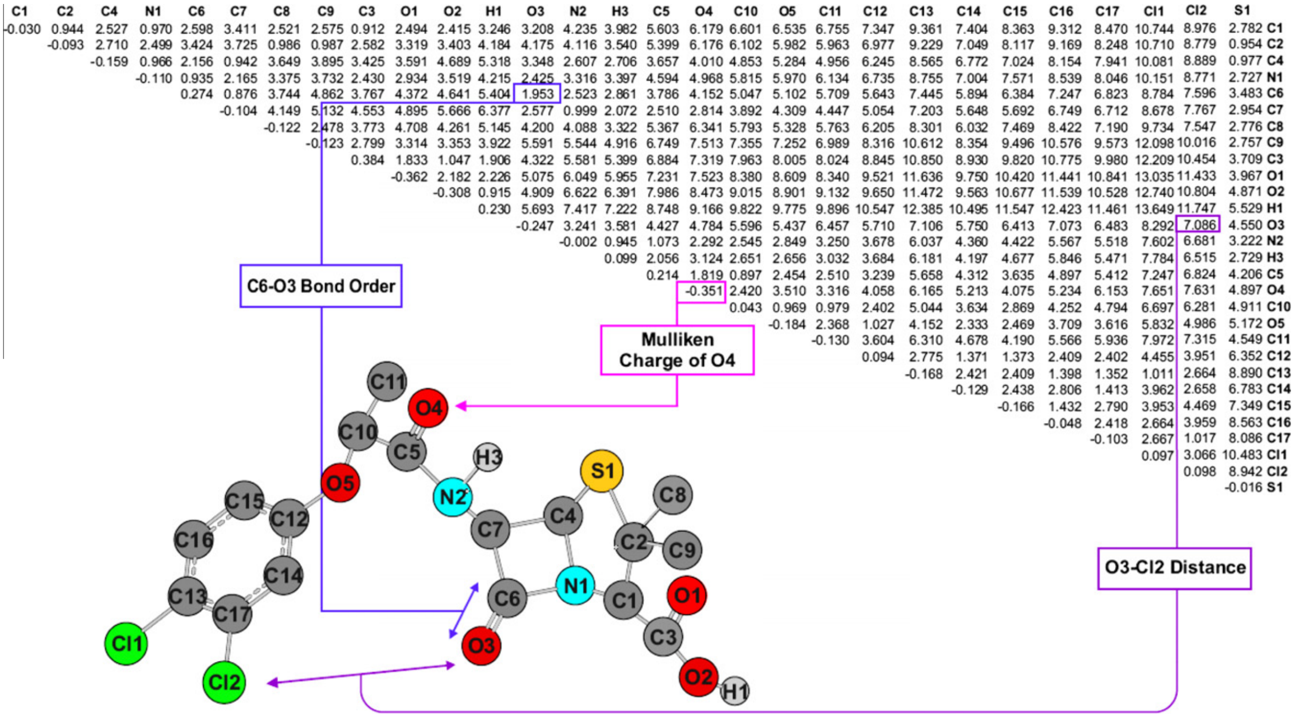
Combining different methods in a hybrid method, it is possible to minimize the errors and take advantage of the good features of each method. In this work, a hybrid 4D-QSAR approach (EC–GA) that combines the electron conformational (EC) and genetic algorithm optimization (GA) methods was applied in order to explain the pharmacophore (Pha) and to predict antibacterial activity by studying 87 compounds in the class of β-lactam antibiotics (known as penicillin) derivatives. This method is based on the generation of a conformational ensemble profile for each compound instead of only one conformation, followed by the calculation of descriptors for a set of compounds. The fourth dimension refers to the possibility of representing each ligand molecule as an ensemble of conformations and orientations, thereby reducing the tendency in identifying the bioactive conformation and orientation (4D-QSAR). In our previous studies, this method was successfully performed as a 4D-QSAR procedure to identify the pharmacophore for benzotriazines and triaminotriazine derivatives, and quantitative prediction of activity.[27,28]

2. Materials and methods

The EC method is aimed at establishing rules for the prediction of different activities, based on the pharmacophores found previously by specific EC calculations.[15,21] The method uses data obtained from quantum-chemical calculations and arranged in the form of electron-conformational matrices of conjunction (ECMC), which combine geometric and electronic features.[18] Bersuker and Dimoglo presented the so-called electron conformational (EC) method of Pha identification and quantitative bioactivity prediction in drug design, which is basically different from ETM.[29] This method analyzes molecules represented by matrices known as electronic-topological matrices of congruity for only one conformer of the molecule.[30–33]

The geometries and electronic structure for penicillin derivatives were optimized with the parametric model number 3 (PM3) method using analytical gradients. The purpose of conformational analysis is to obtain a description of the three-dimensional structure of molecules. Such knowledge is required in order to understand the interactions between molecules. If the energies of different conformers are known it should be possible to calculate their relative abundances by Boltzmann weighting. Heavily populated conformations mean those with energies <1.5 kcal/mol above the ground state conformation. Since conformers of low energy have a larger population than other conformers according to Boltzmann distribution, these conformers are more responsible than others.[15] Then we calculate the electronic structure of each of these conformations and arrange the corresponding electronic and geometrical parameters in a matrix $n \times n$ (n is the number of atoms), called the ECMC (Fig. 1). The ECMC, which is specific ECM language for compound structure description, is a square matrix that is symmetric with respect to the diagonal elements. Hence, only the upper part of each ECMC is kept in the memory of the computer and processed by EMRE software which was written by our research group based on the DELPHI program[34] and accepts various standard structure formats as input: spartan.txt or gaussian.out file. The diagonal elements a_{ii} in the ECMC, where i represents the ith atom in the molecule, are one of the electronic atomic characteristics (local atomic characteristics) such as atomic charges and valence activities. The off-diagonal elements a_{ij} are of two kinds, one of which is for chemical bonds and the other is for chemically non-bonded atoms. If i and j label chemically bonded atoms, then a_{ij} is one of the electronic parameters of the i–j bond such as bond order. As a bond property, bond order is often chosen because it reflects bond strength and, together with atomic charges and 3D distances (in Å), it gives valuable information on the electron density distribution in the 3D space of a molecule.[29] If i and j label non-bonded atoms, then a_{ij} is the interatomic distance between the ith and jth atoms (R_{ij}). Under fixed atomic and bond parameters that are deemed most important for activity demonstration, the ECMC of each conformer under consideration is formed.[31,32] The compounds under study (87 molecules) are shown in Table 1. The ECMC of the lowest energy conformer of the most active compound (compound **73**) in the penicillin series is shown in Figure 1.

The pharmacophore is commonly defined as an arrangement of molecular features or fragments forming a necessary but not sufficient condition for biological activity.[35,36] A three-dimensional pharmacophore is defined by a critical geometric arrangement of such features or fragments.[37] To begin the identification of pharmacophore groups, the compounds with percentage fraction bound (PFB) $\geqslant$ 80.00 were classified as high activity compounds (46 compounds) and molecules with PFB <80.00 were considered to be low activity compounds (41 compounds). To find pharmacophores, a template active compound and the rest of the compound set are compared as weighted graphs. For each compound (high active or low active) taken as a template, its ECMC was compared with the ECMCs of the rest of the compounds in both classes of the series under study within tolerances.[32] Flexibility limits are quite important for the realization of the Pha. For example, further growth of the upper limit of any elements of the submatrix causes weakly active compounds to include the feature responsible for high activity.[15,38] The comparison resulted in a few common



C1	C2	C4	N1	C6	C7	C8	C9	C3	O1	O2	H1	O3	N2	H3	C5	O4	C10	O5	C11	C12	C13	C14	C15	C16	C17	Cl1	Cl2	S1	
-0.030	0.944	2.527	0.970	2.598	3.411	2.521	2.575	0.912	2.494	2.415	3.246	3.208	4.235	3.982	5.603	6.179	6.601	6.535	6.755	7.347	9.361	7.404	8.363	9.312	8.470	10.744	8.976	2.782	C1
	-0.093	2.710	2.499	3.424	3.725	0.986	0.987	2.582	3.319	3.403	4.184	4.175	4.116	3.540	5.399	6.176	6.102	5.982	5.963	6.977	9.229	7.049	8.117	9.169	8.248	10.710	8.779	0.954	C2
		-0.159	0.966	2.156	0.942	3.649	3.895	3.425	3.591	4.689	5.318	3.348	2.607	2.706	3.657	4.010	4.853	5.284	4.956	6.245	8.565	6.772	7.024	8.154	7.941	10.081	8.889	0.977	C4
			-0.110	0.935	2.165	3.375	3.732	2.430	2.934	3.519	4.215	2.425	3.316	3.397	4.594	4.968	5.815	5.970	6.134	6.735	8.755	7.004	7.571	8.539	8.046	10.151	8.771	2.727	N1
				0.274	0.876	3.744	4.862	3.767	4.372	4.641	5.404	1.953	2.523	2.861	3.786	4.152	5.047	5.102	5.709	5.643	7.445	5.894	6.384	7.247	6.823	8.784	7.596	3.483	C6
					-0.104	4.149	5.132	4.553	4.895	5.666	6.377	2.577	0.999	2.072	2.510	2.814	3.892	4.309	4.447	5.054	7.203	5.648	5.692	6.749	6.712	8.678	7.767	2.954	C7
						-0.122	2.478	3.773	4.708	4.261	5.145	4.200	4.088	3.322	5.367	6.341	5.793	5.328	5.763	6.205	8.301	6.032	7.469	8.422	7.190	9.734	7.547	2.776	C8
							-0.123	2.799	3.314	3.353	3.922	5.591	5.544	4.916	6.749	7.513	7.355	7.252	6.989	8.316	10.612	8.354	9.496	10.576	9.573	12.098	10.016	2.757	C9
								0.384	1.833	1.047	1.906	4.322	5.581	5.399	6.884	7.319	7.963	8.005	8.024	8.845	10.850	8.930	9.820	10.775	9.980	12.209	10.454	3.709	C3
									-0.362	2.182	2.226	5.075	6.049	5.955	7.231	7.523	8.380	8.609	8.340	9.521	11.636	9.750	10.420	11.441	10.841	13.035	11.433	3.967	O1
										-0.308	0.915	4.909	6.622	6.391	7.986	8.473	9.015	8.901	9.132	9.650	11.472	9.563	10.677	11.539	10.528	12.740	10.804	4.871	O2
											0.230	5.693	7.417	7.222	8.748	9.166	9.822	9.775	9.896	10.547	12.385	10.495	11.547	12.423	11.461	13.649	11.747	5.529	H1
												-0.247	3.241	3.581	4.427	4.784	5.596	5.437	6.457	5.710	7.106	5.750	6.413	7.073	6.483	8.292	7.086	4.550	O3
													-0.002	0.945	1.073	2.292	2.545	2.849	3.250	3.678	6.037	4.360	4.422	5.567	5.518	7.602	6.681	3.222	N2
														0.099	2.056	3.124	2.651	2.656	3.032	3.684	6.181	4.197	4.677	5.846	5.471	7.784	6.515	2.729	H3
															0.214	1.819	0.897	2.454	2.510	3.239	5.658	4.312	3.635	4.897	5.412	7.247	6.824	4.206	C5
																-0.351	2.420	3.510	3.316	4.058	6.165	5.213	4.075	5.234	6.153	7.651	7.631	4.897	O4
																	0.043	0.969	0.979	2.402	5.044	3.634	2.869	4.252	4.794	6.697	6.281	4.911	C10
																		-0.184	2.368	1.027	4.152	2.333	2.469	3.709	3.616	5.832	4.986	5.172	O5
																			-0.130	3.604	6.310	4.678	4.190	5.566	5.936	7.972	7.315	4.549	C11
																				0.094	2.775	1.371	1.373	2.409	2.402	4.455	3.951	6.352	C12
																					-0.168	2.421	2.409	1.398	1.352	1.011	2.664	8.890	C13
																						-0.129	2.438	2.806	1.413	3.962	2.658	6.783	C14
																							-0.166	1.432	2.790	3.953	4.469	7.349	C15
																								-0.048	2.418	2.664	3.959	8.563	C16
																									-0.103	2.667	1.017	8.086	C17
																										0.097	3.066	10.483	Cl1
																											0.098	8.942	Cl2
																												-0.016	S1

Figure 1. ECMC of the lowest energy conformer of compound 73 consisting of 29-atoms without hydrogen (due to its symmetry, only upper triangle is shown). Electronic and geometric values of carbon-hydrogen bond are not taken since in each molecule it has an equivalent effect.

structural fragments for high and low activity compounds. The fragments were found as sub-matrices of the ECMCs corresponding to templates (they will be referred to as electron conformational sub-matrices of activity, or ECSAs).

A criterion that is commonly used in structural methods for evaluating the probability of an activity fragment Pha occurrence in a series under study is given by the following formulas:[30]

$$P_\alpha = (n_1 + 1)/(n_1 + n_3 + 2);$$

$$\alpha_a = (n_1 * n_4 - n_2 * n_3)/(m_1 * m_2 * m_3 * m_4)^{1/2}$$

where n_1 and n_2 are the numbers of molecules possessing and not possessing, respectively, the feature of activity (predicted by the ECM) in the class of compounds; n_3 and n_4 have analogous meaning in the weak active compounds; and m_1 and m_2 are the numbers of molecules in the class of active and weak active compounds $m_3 = n_1 + n_3$; $m_4 = n_2 + n_4$. In this way, P_α evaluates the deposit of only active molecules, while α_a reflects the deposit of both active and low active compounds in the feature of the activity found.[39] Then, without setting any constraints on tolerance values, maximum tolerance values are calculated for all conformers of all compounds.

The strategy of QSAR modeling is to condense the relationship between the structure of molecules and their properties into a mathematical expression. During the development of the 4D-QSAR approach one important task is to find the best activity prediction function. The contributions of different conformations of the same compound are taken into account by means of Boltzmann distribution. For the purpose of averaging descriptors, we assigned weights to each of the conformers based on their probability of existence per the Boltzmann distribution. The following general formula of activity of the n-th compound results from the work done by Bersuker et al.:[15]

$$A_n = A_l \frac{\sum_{i=1}^{m_l} e^{\frac{-E_{li}}{RT}}}{\sum_{i=1}^{m_n} e^{\frac{-E_{ni}}{RT}}} \frac{\sum_{i=1}^{m_n} \delta_{ni}[\mathrm{Pha}] e^{-S_{ni}} e^{\frac{-E_{ni}}{RT}}}{\sum_{i=1}^{m_l} \delta_{li}[\mathrm{Pha}] e^{-S_{li}} e^{\frac{-E_{li}}{RT}}} \qquad (1)$$

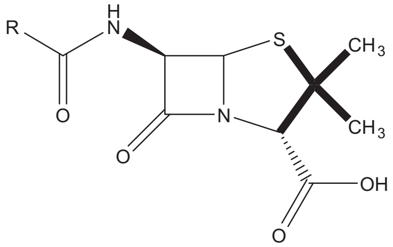
where δ is the Kronecker delta. This is a function of two variables which is 1 if the pharmacophore is present and 0 if not. A_n and A_l are the activities of the nth compound and the reference molecule, and m_n is the number of conformations of the nth molecule. E_{li} is the relative energy of the ith conformation of the reference compound, E_{ni} is the relative energy of the ith conformation of the nth compound (in kcal mol^{-1}), R (kcal mol^{-1} K^{-1}) is the gas constant, and T is the temperature in Kelvin. In all of the molecular systems under consideration that have a Pha, there exist anti-pharmacophore shielding (APS) and auxiliary groups (AG) that affect their activity one way or another. This is the reason why, among the compounds that contain a pharmacophore, some are more active than others. Both the AG and the APS can then be described by means of molecular parameters. The effect of these parameters is determined by introducing the function S to be the sum of all these effects as follows:

$$Sn_i = \sum_{j=1}^{N} \kappa_j a_{ni}^{(j)} \qquad (2)$$

where $a_{ni}^{(j)}$ are the parameters that describe the jth kind of influence in the ith conformation of the nth molecule, and N is the number of chosen parameters. The activity depends exponentially on S ($A \sim e^{-S}$); in this way S is deemed to take into account the specific features of the drug–receptor interaction that determine the activity quantitatively. Using the function S, and taking into account the Boltzmann population of each conformation as a function of its energy and temperature κ_j, the variational (adjustable) constants are calculated. The lsqnonlin function within the statistics toolbox in MATLAB[40] was used to obtain κ_j values of the corresponding model parameters by solving numerically the system of differential equations for the best subset of variables.

The main problem in the quantitative estimate of the activity after identification of the Pha is to choose the N parameters (variables) $a_{ni}^{(j)}$ in Eq. 2 and to determine their weights, namely the κ_j constants. To this end, the estimated values of the parameters for the active conformation of the compounds in the training set were

Table 1
Molecular structures, conformer numbers, experimental and predicted activity values of penicillin series for 12 parameters. Model 1 and 2 based on ensemble of conformers and single conformer, respectively



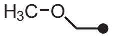
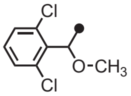
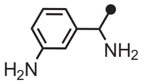
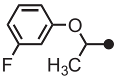
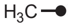
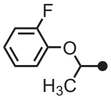
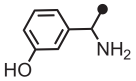
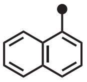
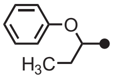
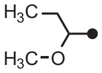
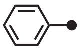
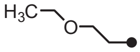
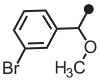
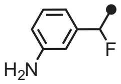
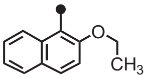
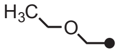
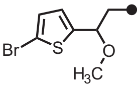
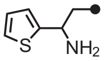
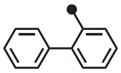
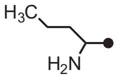
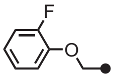
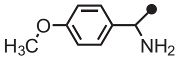
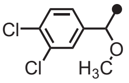
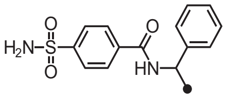
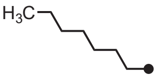
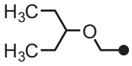
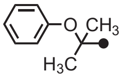
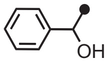
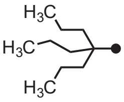
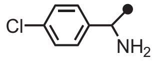
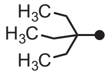
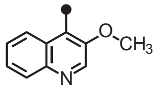
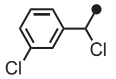
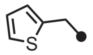
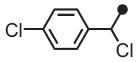
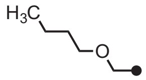
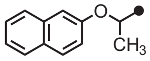
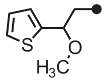
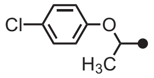
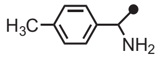
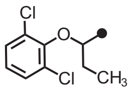
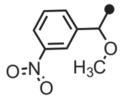
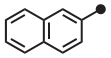
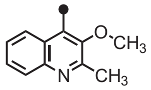
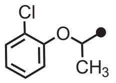
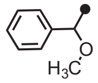
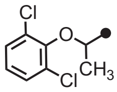
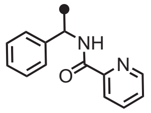
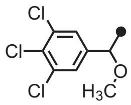
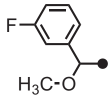
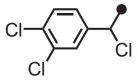
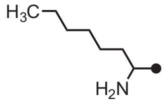
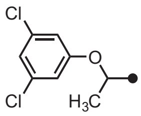
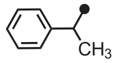
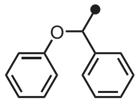
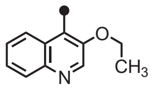
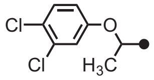
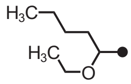
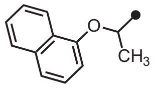
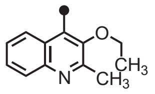
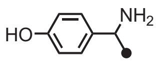
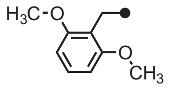
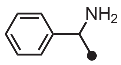
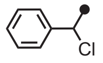
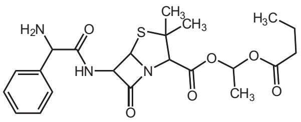
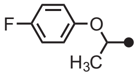
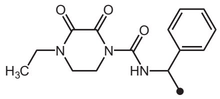
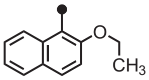
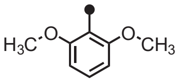
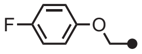
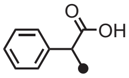
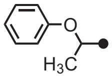
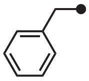
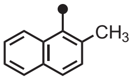
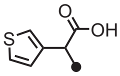
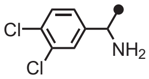
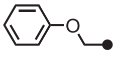
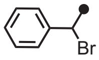
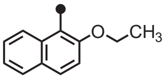
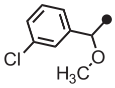
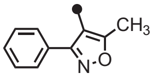
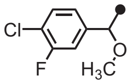
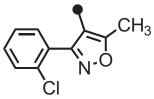
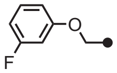
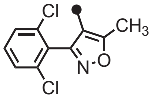
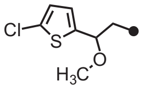
ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]		ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]	
				Model 1	Model 2					Model 1	Model 2
1		14	7.200	13.976	14.536	**45**		12	84.000	88.719	84.508
2		13	12.000	15.839	14.843	**46**		14	84.000	85.726	84.496
3		8	15.000	12.116	12.848	**47**		26	86.000	85.289	81.187
4		20	16.800	20.584	20.884	**48**		9	86.000	78.787	61.333
5	R all values = 0	7	18.000	11.208	10.904	**49**		13	86.100	76.053	68.050
6		13	20.000	19.631	20.825	**50**		8	87.000	66.091	50.969
7		15	25.000	22.472	24.147	**51**		9	88.000	98.080	94.185
8		22	26.000	18.928	19.248	**52**		8	89.300	91.148	79.952
9		9	28.000	15.760	17.650	**53**		11	89.700	103.736	93.863
10		12	32.000	22.806	21.471	**54**		9	91.000	97.516	101.115
11		11	33.000	14.959	15.690	**55**		10	91.500	104.171	88.240
12		18	38.000	21.064	25.766	**56**		10	92.000	103.188	99.505
13		18	42.000	49.931	55.249	**57**		13	92.400	91.534	83.098
14		15	47.000	51.131	51.453	**58**		12	92.500	103.867	102.740
15		10	53.200	62.286	58.569	**59**		15	93.300	113.108	123.777

Table 1 (*continued*)

ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]		ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]	
				Model 1	Model 2					Model 1	Model 2
16		14	55.000	48.351	45.921	**60**		20	93.600	85.512	93.056
17		24	57.000	52.031	58.251	**61**		24	94.000	99.412	110.434
18		25	58.000	35.940	34.456	**62**		19	94.000	101.361	107.806
19		19	58.800	43.766	38.683	**63**		9	94.700	83.411	69.622
20		16	59.000	43.572	43.460	**64**		8	94.800	76.341	78.278
21		11	60.000	48.663	38.610	**65**		12	95.200	96.056	89.622
22		17	60.000	60.351	56.591	**66**		13	95.600	96.300	79.976
23		31	61.700	44.572	41.741	**67**		11	96.000	85.054	69.116
24		8	62.000	35.615	37.135	**68**		12	96.000	92.335	73.518
25		14	63.000	57.302	54.341	**69**		14	96.500	105.463	111.317
26		13	65.000	53.953	60.073	**70**		21	97.000	84.679	77.822
27		9	66.200	54.331	55.481	**71**		13	97.000	112.220	96.920
28		17	68.000	75.042	87.817	**72**		9	97.200	80.086	85.481
29		21	69.700	69.074	80.442	**73**		15	97.400	97.400	97.400
30		17	74.000	62.615	72.121	**74**		8	97.400	84.436	67.699
31		14	74.500	91.548	90.003	**75***		14	18.000	41.752	32.896
32		13	77.000	60.485	70.574	**76***		15	18.000	35.063	25.814
33		10	78.000	65.667	70.117	**77***		8	20.000	32.147	39.031

(*continued on next page*)

Table 1 (*continued*)

ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]		ID[a]	R	C_n[b]	PFB_{exp}[c]	PFB_{pred}[d]	
				Model 1	Model 2					Model 1	Model 2
34		10	80.000	94.607	85.251	**78***		14	30.000	40.857	39.655
35		26	80.000	81.474	84.710	**79***		18	39.000	42.896	35.406
36		16	81.500	69.729	59.368	**80***		3	50.000	42.895	49.039
37		13	81.500	75.053	75.642	**81***		13	60.000	60.919	52.278
38		11	82.100	94.490	88.083	**82***		18	65.000	59.107	50.545
39		17	82.200	62.418	56.271	**83***		11	80.000	57.600	49.049
40		11	82.500	80.847	78.331	**84***		4	89.000	80.849	72.463
41		25	83.000	87.675	67.853	**85***		9	92.000	68.756	56.154
42		32	83.000	77.486	89.589	**86***		15	95.000	73.853	68.109
43		24	83.500	66.842	64.630	**87***		12	96.000	80.302	71.450
44		12	83.600	69.442	66.140						

[a] ID: Number of compounds. The test set compounds are labeled with an '*' symbol.

[b] C_n: Conformer number of compounds within 1.5 kcal.mol^{-1} of their respective lowest conformations.

[c] PFB_{exp}: experimental percentage fraction bound activity.

[d] PFB_{pred}: predicted percentage fraction bound activity. Model 1and 2 refer to the ensemble of conformers and single conformer, respectively.

used to perform a least-square minimization of the expression $\sum_n |A_n^{exp} - A_n^{pred}|^2$ as a function of the unknown constants κ_j with A_n^{pred} from Eq. 1 and A_n^{exp} from the experimental data. This procedure was carried out using Matlab software in conjunction with the optimization function lsqnonlin, which is a general non-linear least squares fitting algorithm, to fit the data. The numbers 'κ_j', $j = 1, 2,\ldots,N$, obtained in this way characterize the weights of each kind of the $a_{ni}^{(j)}$ parameters in the overall APS/AG influence.[15] In the case of large numbers of (experimental or theoretical) molecular descriptors, the correct selection of relevant descriptors for the QSAR models becomes important.

In this study, two different models were generated by using multiple conformers in model 1 and only one conformation of each compound in model 2. So that not only the lowest energy conformation, but also all the reasonable conformers given in Table 1, enter the final activity formula (Eq. 1) in order to predict the activity. To reduce the large amount of computation time associated with number of conformers, parallel GA, which is a very effective method to handle computationally-expensive problems, was used to solve this problem and to find the best solution. The genetic algorithm (GA) has been widely used for feature optimization in QSAR models.[24] The purpose of variable selection is to select the variables significantly contributing to prediction and discard other variables by a fitness function. To examine the correlation between the fitness values and various subsets in the different number of variables (a_n), it is essential to run the GA procedure many times with some values (in this work a_n values = 1–13). In our study, the values of empirical parameters necessary for the GA computation are as follows: The number of population, generation and iteration were set at 400, 400 and 500, respectively. The probability of crossover and mutation were set at 0.85 and 0.015, respectively. These values were determined to be optimal after several GA computations with changing the values of empirical parameters.

The fitness function has a great effect on the convergence speed of a GA process. We used the predictive residual sum of squares (PRESS) as a fitness function (Eq. 3). In this study, the fitness value for each chromosome was calculated by leave-one-out cross-validation (LOO-CV). The PRESS is a standard index to measure the accuracy of a modeling method based on the LOO cross-validation technique for a number of available examples n. PRESS is defined as the sum of the squared difference between predicted (pred) and experimental (exp) values and can be written as:

$$\mathrm{PRESS}_N = \sum_{n=1}^{N} \left| A_n^{\mathrm{exp}} - A_n^{\mathrm{pred}} \right|^2 \tag{3}$$

where A_n^{exp} are the experimental activities and A_n^{pred} are the predicted activities in the LOO cross-validation model, with N parameters being used (in this case $N = 12$). The LOO technique is a good way to quantitatively evaluate the predictive ability and robustness of a model by predicting each compound's activity using a QSAR model built based on information of the remaining compounds, which avoids the effect of a compound on its own activity prediction. A cross-validated correlation coefficient (q^2) was used to measure the predictability of the model and the conventional correlation coefficient (R^2) was used to measure the quality of the model. The cross-validated correlation coefficient q^2, as a measure of the prediction performance of the model can then be written as:[41]

$$q^2 = 1 - \frac{\sum_{n=1}^{N} \left| A_n^{\mathrm{exp}} - A_n^{\mathrm{pred}} \right|^2}{\sum_{n=1}^{N} \left| A_n^{\mathrm{exp}} - \overline{A}_n^{\mathrm{exp}} \right|^2} \equiv 1 - \frac{\mathrm{PRESS}}{\mathrm{SSY}} \tag{4}$$

where N is the total number of training compounds in the entire data set; SSY is the sum of the squares of deviations of the experimental values (A_n^{exp}) from their mean; and A_n^{exp} and $\overline{A}_n^{\mathrm{exp}}$, respectively, are the measured and averaged (over the entire training set) values of the dependent variable. The smaller the PRESS is, the better the predictability of the model. When its value is than SSY this shows that the model predicts better than chance and can be considered statistically significant. SSY is the sum of squares associated with the corresponding sources of variation. External validation refers to a validation exercise in which the chemical structures selected for inclusion in the test set are different from those included in the training set, but which should be representative of the same chemical domain. The QSAR model developed by using the training set chemicals was then applied to the test set chemicals in order to verify the predictive ability of the model. External validation is the only way to determine the true predictive power of a QSAR model.[42] Two different expressions for the calculation of q^2 from an external evaluation set were discussed by Schuurmann et al.[41,43]

These expressions are:

$$q^2_{\mathrm{ext1}} = 1 - \frac{\sum_{n=1}^{N} \left| A_{n_{\mathrm{test}}}^{\mathrm{exp}} - A_{n_{\mathrm{test}}}^{\mathrm{pred}} \right|^2}{\sum_{n=1}^{N} \left| A_{n_{\mathrm{test}}}^{\mathrm{exp}} - \overline{A}_{\mathrm{training}}^{\mathrm{exp}} \right|^2} \tag{5}$$

$$q^2_{\mathrm{ext2}} = 1 - \frac{\sum_{n=1}^{N} \left| A_{n_{\mathrm{test}}}^{\mathrm{exp}} - A_{n_{\mathrm{test}}}^{\mathrm{pred}} \right|^2}{\sum_{n=1}^{N} \left| A_{n_{\mathrm{test}}}^{\mathrm{exp}} - \overline{A}_{n_{\mathrm{test}}}^{\mathrm{exp}} \right|^2} \tag{6}$$

where N = number of tested molecules, A_n^{exp} = experimental activity, A_n^{pred} = predicted activity without using the left-out compound in the model building and $\overline{A}_n^{\mathrm{exp}}$ = average of experimental activities.

3. Results and discussion

4D-QSAR analysis was applied to a series of 87 (a training set of 74 and a test set of 13, commercial penicillins) penicillin inhibitors of human serum protein, with a PFB ranging from 7.2 to 97.4. The chemical structures and experimental activities were obtained from the literature.[6]

EC–GA is a sophisticated hybrid approach that combines the EC method of pharmacophore identification and bioactivity prediction with GA as a powerful optimization. The compounds under study, along with their common structural skeletons and conformer number of compounds for both the training and test compounds, and experimental ($\mathrm{PFB_{exp}}$) and predicted activity ($\mathrm{PFB_{pred}}$) values, which were obtained by using the ensemble of conformers (model 1) and single conformer (model 2), are shown in Table 1.

Semi-empirical quantum chemical calculation at PM3 level was used to find the optimum 3D conformers' geometry of the studied molecules. ECMCs were constructed from conformational analysis data and the electronic structure calculation of each of the molecules in the compound series by the EMRE programme. The electronic and geometric values of the carbon–hydrogen bond were not taken since in each molecule it has an equivalent effect (Fig. 1). We chose the ECMC of the lowest energy conformer of compound with the highest activity as a template (compound **73**) and compared it with the ECMCs of conformers with the lowest energy of other compounds within tolerances.[15] By gradually changing the limits of tolerance for diagonal elements (charges) and non diagonal elements (bond orders or interatomic distances), we finally obtained the tolerance limits indicated in the ECSA in Table 2, which give the best separation of the active compounds from the similar inactive or low active ones. The first submatrix in Table 2 shows the ECSA of the reference compound; the second submatrix corresponds to the tolerance values for compounds with high activity; and the third submatrix shows tolerance values for compounds with low activity. Then, without setting any constraints on the tolerance values of the pharmacophore group, the maximum tolerance values were calculated for all conformers of all compounds. The last submatrix shows the tolerance values for the conformers (1212) of all compounds (87). After screening the 46 active compounds within the initial ECSA tolerances not exceeding 0.25 for diagonal and 1.30 for off-diagonal elements, we found that the ECMC submatrix that is common for all of the active molecules contains seven atoms corresponding to N1, C6, C3, O1, H1, O3 and O4 in all of the compounds. N1, C6, C3, O1, H1 and O3 atoms belong to the 6-aminopenicillanic acid (6-APA) group.

We found that out of the seven pharmacophore atoms four negatively charged, while three of them are positively charged. The C6, N1 and O3 form a rigid plane, while the positions of C3, O1, H1 and O4 atoms are fairly flexible. The account of this flexibility within the Pha geometry and its influence on the value of activity is a special feature of the EC method. We see that the tolerance matrix for less active compounds does not fit to the tolerance matrix for higher active compounds (Table 2). Tolerance values existing in compounds of high activity values are usually lower than those existing in low activity compounds (Table 2). For example, tolerance values of distance between the O4 and H1 atoms for higher and less active compounds are ±0.777 and ±2.732, respectively. n_1, n_2, n_3 and n_4 values are found to be 46, 0, 5 and 36, respectively. In this way, the parameters expressing the probability of feature realization are high enough $P_\alpha = 0.887$, $\alpha_a = 0.890$.

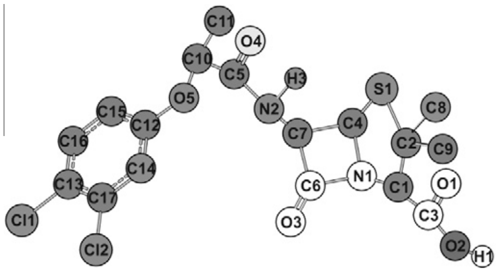
For compounds with a Pha we predict bioactivity values using Eq (2). Eighty-seven molecules were divided into a training set of 74 and a test set of 13 commercial penicillin compounds, which were used to validate the QSAR models. The test compounds were not included in model generation. The most active molecule **73** was used as a template molecule for alignment.

In the QSAR approach, molecules can be represented by a wide variety of (theoretical) molecular descriptors, which are used as independent variables in the model. A total of 400 molecular descriptors, belonging to different descriptor families, were calculated for 1212 conformers of the 87 compounds, using EMRE software; this program has an effective descriptor generation based on the information given by the SPARTAN 06[44] output file for molecular structures.

The EMRE program can extract and calculate about 1000 molecular descriptors, including geometrical, quantum-chemical,

Table 2

(a) ECSA of reference compound (**73**) for penicillin derivatives b) Tolerance matrix of ECSA values for 46 compounds with high activity, (c) Tolerance values for 41 compounds with low activity, (d) Tolerance values for 1212 conformations of 87 compounds



N1	C6	C3	O1	H1	O3	O4	Pha atoms
(a) ECSA (Pha) of reference compound							
−0.110	0.935	2.430	2.934	4.215	2.425	4.968	**N1**
	0.274	3.767	4.372	5.404	1.953	4.152	**C6**
		0.384	1.833	1.906	4.322	7.319	**C3**
			−0.362	2.226	5.075	7.523	**O1**
				0.230	5.693	9.166	**H1**
					−0.247	4.784	**O3**
						−0.351	**O4**
(b) Tolerance values for 46 compounds with high activity							
±0.061	±0.035	±0.025	±0.580	±0.498	±0.006	±0.578	**N1**
	±0.036	±0.038	±0.298	±0.284	±0.031	±0.617	**C6**
		±0.002	±0.025	±0.016	±0.040	±0.435	**C3**
			±0.027	±0.034	±0.128	±1.275	**O1**
				±0.003	±0.087	±0.777	**H1**
					±0.016	±0.983	**O3**
						±0.036	**O4**
(c) Tolerance values for 41 compounds with low activity compounds							
±0.037	±0.020	±0.046	±0.660	±1.796	±0.003	±0.562	**N1**
	±0.022	±0.146	±0.319	±1.756	±0.018	±0.617	**C6**
		±0.022	±0.051	±0.524	±0.347	±0.421	**C3**
			±0.063	±1.190	±0.361	±1.262	**O1**
				±0.027	±1.486	±2.732	**H1**
					±0.015	±1.040	**O3**
						±0.023	**O4**
(d) Tolerance values for 1212 conformations of 87 compounds							
±0.064	±0.041	±0.050	±0.671	±1.811	±0.009	±0.578	**N1**
	±0.042	±0.166	±0.353	±1.805	±0.046	±0.624	**C6**
		±0.024	±0.051	±0.534	±0.384	±0.459	**C3**
			±0.063	±1.192	±0.396	±1.311	**O1**
				±0.027	±1.582	±2.732	**H1**
					±0.027	±1.040	**O3**
						±0.041	**O4**

electrostatic and thermodynamic descriptors. Several hundreds, even thousands, of descriptors can be generated in QSAR studies. To avoid the danger of over-fitting and to make a stable model, only a subset of available independent variables should be selected in QSAR analysis. For this reason, we applied the genetic algorithm (GA) procedure to select the variables. The genetic algorithm codes in this study were written in Matlab and were run on parallel (multi-core and multi-processor) computers. The variational constant, κ_j, in Eq. 2 was mathematically optimized using the Matlab toolbox function lsqnonlin.[40]

Various 4D-QSAR models were generated for the study and the best was selected on the basis of the statistically significant parameters obtained. The predictive power of the 4D-QSAR models, derived using the training set, was assessed by predicting the biological activity of the test set molecules. Since the optimum number of variables is not known in advance, several runs were needed to examine the relationship between the predictive power of a model (q^2) and the number of descriptors selected. The plot of squared correlation coefficients versus parameter numbers is shown in Figure 2. The amount of descriptors was in the range of 1–13 for the training and test sets. According to the q^2 values, results indicated that between 5 and 13 parameters can acceptable. Statistical results of the model 1 and 2 generated by the EC–GA method for penicillin derivatives are tabulated in Table 3. The optimum 12 molecular parameters, selected with GA and κ_j values used in activity calculation for penicillin derivatives, were shown for both models in Table 4. κ_j values were different in both models because of reoptimization for model 2.

In Table 4, $a^{(1)}$ is softness related to high polarizability and low electronegativity. The softness used for the intermolecular reactivity trend, S, is simply the inverse of the hardness, $S = 1/\eta(2/\varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}})$.[45] $a^{(2)}$ is the electronegativity of a functional group related to both its hydrophobic property and its ability to form hydrogen bonds with surrounding (bioreceptor) molecules, and it is usually considered a very important factor for describing biological system properties.[46] Electronegativity = $1/2(\varepsilon_{\text{LUMO}} + \varepsilon_{\text{HOMO}})$. $a^{(3)}$ is the nucleophilicity index. Parr et al. introduced the global electrophilicity index (ω).[47] The electrophilicity index measures energy stabilization when the energy of a ligand is reduced due to optimal electron flow between donor and acceptor. On the basis of the assumption that electrophilicity and nucleophilicity are inversely related to each other, Chattaraj et al.[48] suggested a

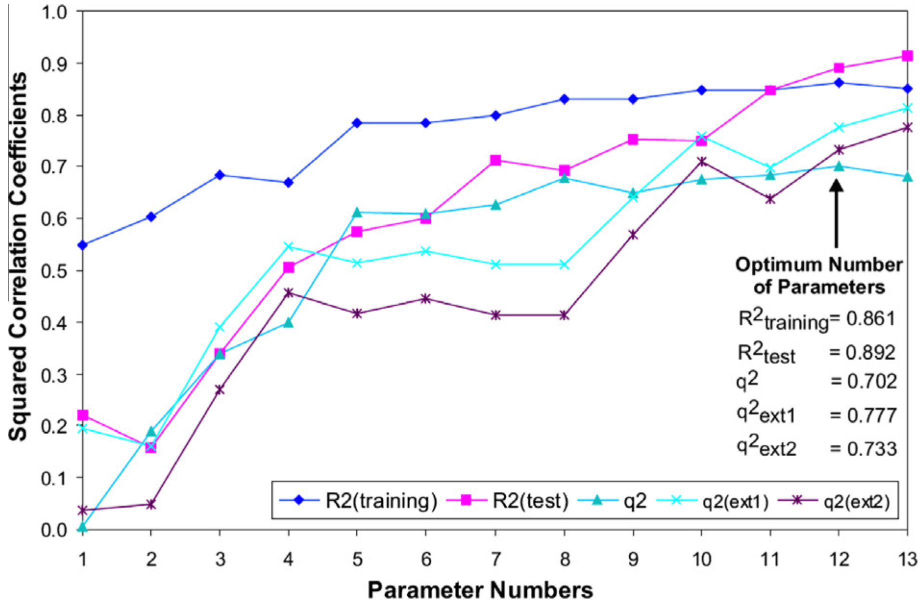


Figure 2. Dependence of regression coefficients in calibration (R^2) and cross-validation correlation coefficients (q^2) on number of semi-empirical chemical descriptors used in model 1.

Table 3
Statistical results of model 1 and 2 generated by EC–GA method for penicillin derivatives

$a^{(j)}$	$R^2_{training}$		R^2_{test}		q^2		q^2_{ext1}		q^2_{ext2}	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
1	0.550	0.104	0.223	0.040	0.005	−2.001	0.196	−1.840	0.037	−2.402
2	0.602	0.601	0.159	0.170	0.189	0.167	0.161	0.167	0.050	0.002
3	0.684	0.675	0.338	0.334	0.339	0.278	0.391	0.357	0.271	0.229
4	0.668	0.660	0.504	0.362	0.400	0.346	0.545	0.424	0.456	0.310
5	0.784	0.642	0.574	0.237	0.612	0.292	0.513	0.234	0.417	0.083
6	0.786	0.643	0.600	0.225	0.611	0.275	0.536	0.222	0.444	0.068
7	0.800	0.752	0.712	0.625	0.627	0.556	0.511	0.428	0.414	0.314
8	0.832	0.771	0.694	0.754	0.678	0.578	0.511	0.543	0.415	0.453
9	0.832	0.766	0.754	0.742	0.650	0.543	0.640	0.595	0.568	0.515
10	0.849	0.781	0.750	0.808	0.676	0.570	0.759	0.592	0.711	0.512
11	0.849	0.762	0.847	0.766	0.683	0.507	0.698	0.605	0.638	0.526
12	0.861	0.774	0.892	0.840	0.702	0.514	0.777	0.641	0.733	0.570
13	0.851	0.776	0.914	0.838	0.681	0.500	0.814	0.638	0.777	0.567

R^2, regression coefficient for training and test set; q^2, internal and external cross-validated correlation coefficients and $a^{(j)}$, parameter number. Model 1and 2 refer to the ensemble of conformers and single conformer, respectively.

Table 4
12 κ_j values and molecular parameters used in activity calculation for model 1 and 2

$a^{(j)}_{ni}$	Molecular parameters	κ_j values	
		Model 1	Model 2
$a^{(1)}$	Softness (eV^{-1})	15.726	4.917
$a^{(2)}$	Electronegativity (eV)	1.886	0.534
$a^{(3)}$	Nucleophilicity index (eV^{-1})	13.530	3.755
$a^{(4)}$	Rotational entropy of molecule (kcal/mol)	−0.285	−0.346
$a^{(5)}$	Log *P* (Partition coefficient)	−0.130	−0.168
$a^{(6)}$	The angle (radian) between line of C2–C8 atoms and C3–O1–O2 plane	7.636	3.118
$a^{(7)}$	Fukui atomic nucleophilic reactivity index of S1 atom	−0.394	−0.266
$a^{(8)}$	Fukui atomic electrophilic reactivity index of N2 atom	−24.825	−18.935
$a^{(9)}$	Fukui atomic electrophilic reactivity index of C6 atom	−40.744	−54.586
$a^{(10)}$	C6–D* distance (Å)	−0.258	−0.066
$a^{(11)}$	O3–D* distance (Å)	0.268	0.102
$a^{(12)}$	CFD (positive ionizable site) basic site number of positive ionizable centers	0.175	0.240

Model 1and 2 refer to the ensemble of conformers and single conformer, respectively.
* The symbol "D" represents the farthest atom in R group, excluding hydrogen as a farthest atom. For example *D* refers to the atom Cl1 for the reference compound (**73**).

multiplicative inverse of the electrophilicity index (1/ω), as well as an additive inverse (1 − ω).[49] One important descriptor found in our model (Table 4) is the rotational entropy ($a^{(4)}$) calculated at a constant temperature of 298.15 K, which is a measure of the rotational degree of freedom of the molecule and can be related to its size, mass distribution and flexibility. Thus, this descriptor's value is mainly influenced by the presence of different substituent groups. When a molecule binds to a protein, it loses a significant amount of rotational entropy. Estimates of the associated energy barrier vary widely in the literature yet accurate estimates are important in the interpretation of results from fragment-based drug discovery techniques.[50] $a^{(5)}$ is log *P*, defined as the logarithm of the partition coefficient between *n*-octanol and water, is an important parameter to judge a molecule's drug likeness. Log *P* has been widely used as a measure of lipophilicity, and it is critical for both the pharmacokinetic and pharmacodynamic behavior of a molecule. It is an important parameter used in QSAR studies of biological activity prediction. $a^{(6)}$ is the angle between line and plane (Fig. 3). $a^{(7)}$ is the Fukui atomic nucleophilic reactivity index. $a^{(8)} - a^{(9)}$ are the Fukui atomic electrophilic reactivity index. Within Fukui's Frontier Molecular Orbital (FMO) theory,[51,52] the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) play fundamental roles in the interpretation of chemical reactivity, particularly toward nucleophiles or electrophiles.[53] Local quantities such as the Fukui function and local softness define the reactivity/selectivity of a specific site in a molecule. The Fukui atomic nucleophilic and electrophilic reactivity index that varies from point to point in an atom may be defined by the following formulas: $N_A = \sum_{i \in A} c^2_{i\mathrm{HOMO}}$, $c_{i\mathrm{HOMO}}$ = Highest occupied molecular orbital coefficients and $N_A = \sum_{i \in A} c^2_{i\mathrm{LUMO}}$, $c_{i\mathrm{LUMO}}$ = Lowest unoccupied molecular orbital coefficients, respectively.[54] Fukui functions were used to identify the most reactive sites for the nucleophilic and electrophilic attack of penicillin derivatives. Both the N2 and the C6 atomic centers were found to be suitable electrophilic reactive sites. The S atom of the thiazolidine group was a more reactive nucleophilic centre. $a^{(10)} - a^{(11)}$ are just corresponding interatomic distances employed to take into account the influence of their limited flexibility on the activity (Fig. 3). $a^{(12)}$ is a positive ionizable site. Chemical Function Descriptors (CFD's) are descriptors given to a molecule in order to characterize or anticipate its chemical behavior or to identify commonality among molecules with different structures.[55] A positive ionizable function, which represents any group that is either positively charged or can become positively charged (through protonation at physiological pH), can thus interact with a negatively charged bioreceptor. The positive ionizable feature in penicillin derivatives corresponds to the amide nitrogen atom (N2).

Activity depends exponentially on S ($A \sim e^{-S}$).[17] If the product of the parameter and κ_j values is positive then it shows the APS group, otherwise (if the product is negative) it shows AG. $a^{(1)}$, $a^{(2)}$, $a^{(3)}$, $a^{(6)}$, $a^{(11)}$ and $a^{(12)}$ *are* APS groups, while $a^{(4)}$, $a^{(5)}$, $a^{(7)}$, $a^{(8)}$, $a^{(9)}$ and $a^{(10)}$ are AG groups. The multiplication of the coefficient κ_j with parameter ($a^{(j)}$) should be dimensionless.[15] The reported fitting and validation parameters had very high values for twelve descriptors expressed above to describe the inhibition activity of β-lactams binding against human serum proteins.

In the EC method, it is presented in the previous papers[15] only the lowest conformation that has pharmacophore enters the formula (Eq. 1) which is thus much simplified. However, it cannot be assumed that the lowest energy conformer determines the biological activity of a compound. Although small molecules may have only a single lowest energy conformation but large and flexible molecules do exists in multiple conformations. There are a number of conformations that satisfy the geometric and electronic structure requirements for the active conformation, therefore in 4D-QSAR approach[12] it becomes necessary to include various conformations of the molecules taking into account the Boltzmann populations and dynamics to understand the effects of the all energetically stable conformers on the biological activity.[56] Efficient methods for predicting biological activity should consider the entire range of probable molecular conformations and determine the activity of each conformer. All the conformers interact independently with a bioreceptor. In this study, not only the lowest energy conformation but also all energetically reasonable conformers given enter the final activity formula (Eq. (2)) in order to predict the bioactivity. The calculation results were given and discussed below. Furthermore, all calculation and optimization procedures to predict the biological activity for penicillins were repreformed by taking into account only one conformer of the compound. Although using the lowest energy conformers provides the shortening time of computation, it does not assure to give better results.

The model 1 had a very good descriptive and predictive performance. The quality of R^2 and q^2, and the small standard error (SE) values confirmed the high predictivity of this model. In order to gain a better understanding of the behavior of the fitted data to the model, the descriptors of the model (Table 5) were inspected in detail. The model 1 gave statistically significant results with R^2_{training}, q^2 and $\mathrm{SE}_{\mathrm{training}}$ values of 0.861, 0.702 and 0.044, respectively. The predictive abilities of this model were also evaluated

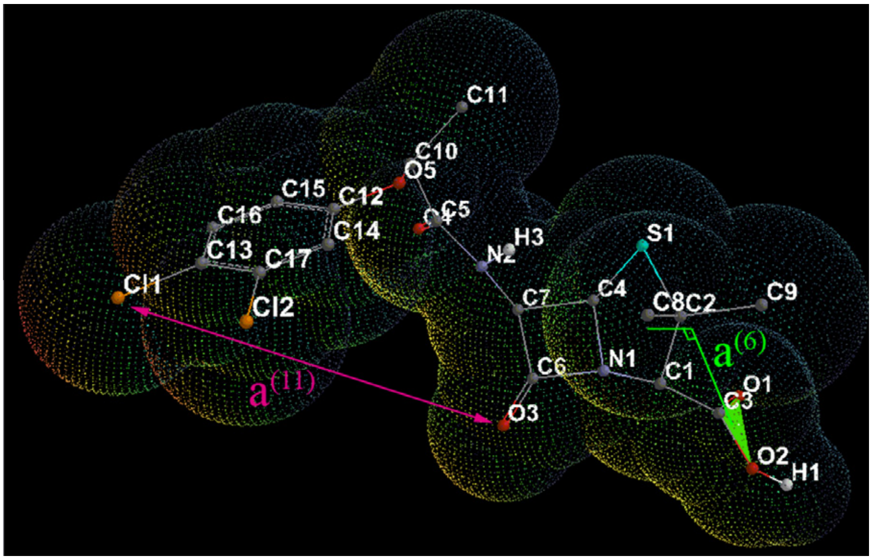


Figure 3. Van der Waals surface of the reference compound. In the figure, $a^{(6)}$ is the angle between line of C2–C8 atoms and C3–O1–O2 plane. $a^{(11)}$ indicates O3–Cl1 distance.

Table 5

E, R^2_{training}, R^2_{test}, q^2, q^2_{ext1} and q^2_{ext2} values showing the contribution of each descriptor to performance of model 1 for PFB activity of penicillin derivatives. q^2_{ext1} and q^2_{ext2} are external validations in the leave-one-out cross-validation for twelve parameters. E statistics allowed us to identify the relative weight each parameter held in the accurate prediction of activities

Eliminated parameters	E	R^2_{training}	R^2_{test}	q^2	q^2_{ext1}	q^2_{ext2}
$a^{(1)}$	0.976	0.862	0.776	0.694	0.712	0.656
$a^{(2)}$	0.859	0.842	0.704	0.653	0.738	0.686
$a^{(3)}$	0.904	0.853	0.746	0.670	0.746	0.696
$a^{(4)}$	0.679	0.820	0.678	0.561	0.586	0.504
$a^{(5)}$	0.533	0.831	0.839	0.440	0.741	0.689
$a^{(6)}$	0.733	0.835	0.784	0.593	0.709	0.651
$a^{(7)}$	0.496	0.736	0.715	0.399	0.275	0.131
$a^{(8)}$	0.885	0.857	0.830	0.663	0.627	0.554
$a^{(9)}$	0.940	0.849	0.847	0.683	0.698	0.638
$a^{(10)}$	0.582	0.839	0.847	0.488	0.760	0.712
$a^{(11)}$	0.449	0.826	0.804	0.336	0.643	0.572
$a^{(12)}$	0.893	0.850	0.793	0.666	0.707	0.648

using test compounds which gave R^2_{test} and SE_{test} values of 0.892 and 0.099, respectively. The developed model also possessed promising predictive ability, as discerned by the testing on the external test set, and could be useful to elucidate the relationship between compound structures and biological activities and to facilitate the design of more potent human serum inhibitors. The plots of predicted activities versus experimental values of predicted activity are shown for the training (74 compounds) and test set (13 compounds) in Figure 4.

Based on the data set and parameters as in the model 1, in which compounds were represented by ensemble of conformers as fourth dimension (4D-QSAR), a new model (model 2) was constructed and optimized using only one conformer of each compound (3D-QSAR). As seen in Table 3, model 2 had lower values of non-cross validated and cross validated regression coefficients ($R^2_{\text{training}} = 0.774$, $\text{SE}_{\text{training}} = 0.056$, $R^2_{\text{test}} = 0.840$, $\text{SE}_{\text{test}} = 0.121$, $q^2 = 0.514$, $q^2_{\text{ext1}} = 0.641$ and $q^2_{\text{ext2}} = 0.570$) than model 1 for both training and test sets. These statistical parameters of the model 1 for 13 parameters were given in Table 3. Model 1 yielded satisfactory statistical results with the cross-validated q^2 value and the non-cross-validated R^2 value. Moreover, our present results also showed that external validation values for model 1 ($q^2_{\text{ext1}} = 0.777$ and $q^2_{\text{ext2}} = 0.733$) were significantly higher than model 2 obtained for the same dataset, indicating that models were statistically significant. Based on the above results, model 1 which has a better predictive ability than model 2 can be accurately used in the design of more potent penicillins.

In this study, the effect of using only the pharmacophore parameters (atomic charges, interatomic distances and bond orders values) was also evaluated for the penicillin series by genetic algorithm optimization method. The resulted models were validated by leave-one out cross-validation procedure to check their predictability and robustness. The model for 74 training compounds of penicillins obtained by pharmacophore parameters gave unsatisfactory statistical results. When the training compounds were decreased, the obtained statistical results had insignificant increase for training and test set. Bersuker calculated the relationship only between the pharmacophore parameters and activity without considering another descriptor, and he found high q^2 values for 17 most active compounds out of 51 training set compounds. But regression results for the 51 training compounds using only the pharmacophore parameters were not reported.[18,22]

To estimate the individual influence of each of the twelve molecular descriptors on activity, the E statistics technique[57,58] was applied to the derived EC–GA model. Each descriptor was neglected once and its influence was evaluated with the remaining eleven descriptors. To see which descriptor contributed the most in a given component, we considered the E, R^2_{training}, R^2_{test}, q^2, q^2_{ext1} and q^2_{ext2} values which are displayed in Table 5. The following formula was employed (Eq. 7):[20]

$$E = \frac{\text{PRESS}_N}{\text{PRESS}_{N-1}} \tag{7}$$

$$\text{PRESS}_{N-1} = \sum_{n=1}^{N-1} \left| A_n^{\text{exp}} - A_n^{\text{pred}} \right|^2 \tag{8}$$

where A_n^{pred} represents the predicted and A_n^{exp} the experimental activities in the LOO cross-validation model, with N parameters being used (in this case $N = 12$).

It can be seen from the table that the q^2 values for $a^{(11)}$ (O3-D distance), $a^{(7)}$ (Fukui atomic nucleophilic reactivity index of S1 Atom), $a^{(5)}$ (log P) and $a^{(10)}$ (C6-D distance) are the lowest; therefore, these four parametres are the most influential. This is supported by the relatively large drop of the q^2 value experiences (from 0.702 to 0.336, 0.399, 0.440, 0.488, respectively). From Table 4, we can conclude that the most significant descriptor, according to the E-statistic result, is O3-D distance (Table 4) which is the most relevant descriptor in the equation. The interatomic distances

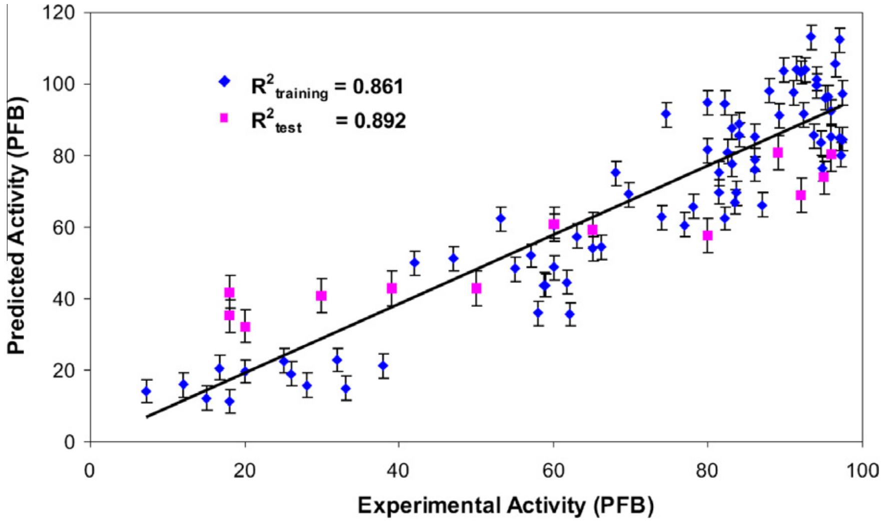


Figure 4. The plots of predicted versus experimental activities are shown by the squared correlation coefficient for training and test sets in model 1.

of both C6-D and O3-D that define the relative spatial dispositions of three significant atoms (the farthest atom of R group, carbonyl carbon, C6, and oxygen, O3, of the lactam ring) may indicate a lipophilic cavity in human serum proteins. In the resulting model in which $a^{(11)}$ is the most influential parameter, the $a^{(1)}$ (softness) parameter can be eliminated without significant loss of accuracy (i.e., with a reduction of q^2 from 0.702 to 0.694).

4. Conclusion

A series of 87 penicillin derivatives were studied to determine the pharmacophore (Pha) and to predict antibacterial activity by means of an original approach called the electron conformational–genetic algorithm method (EC–GA) as a 4D-QSAR. Electron-conformational matrices were effectively used in the search for a system of pharmacophores capable of effective separation of compounds from the examination set into groups of active and low active compounds. Pharmacophores were calculated as sub-matrices containing important spatial and quantum chemistry characteristics. Penicillin activity was controlled by a pharmacophore containing seven atoms with certain electronic and geometrical characteristics corresponding to N1, C6, C3, O1, H1, O3 and O4. Flexibility of molecules was also taken into consideration when searching the fragments of activity (the values of parameters can vary in some limits). In this study, we verified the effectiveness of genetic algorithms as a fast and efficient method to select significant variables to manage complex systems, even when large amounts of descriptors are used as input variables. The model 1 represented by ensemble of conformers also possessed more promising predictive ability than model 2 as discerned by the testing on the external test set, and could be useful to elucidate the relationship between compound structures and biological activities and to facilitate the design of more potent β-lactams binding against human serum protein inhibitors. The predictive ability of the models was validated using a structurally diversified test set of 13 compounds that had not been included in a preliminary training set of 74 compounds. The performances of the model were compared using several statistical measures, including R^2, q^2, and the SE. In this work, a reference compound in a defined conformation was chosen, and all structures in the data set were aligned with a reference within the lowest tolerance values. The 4D-QSAR approach used in this study was a geometric one as the alignment was based on the relative positions of common atoms in space. We presented comprehensive pharmacophore identification, molecular descriptor and activity calculation by a new program (EMRE) which runs on personal computers. This program was developed mainly for computer-aided drug design using four-dimensional quantitative structure–activity relationship methods.

Acknowledgments

This project was financially supported by the Scientific Technical Research Council of Turkey (TUBITAK, Grant No. 107T385). The authors would like to thank Mustafa Yıldırım and Serkan Şahin for their valuable suggestions.

Supplementary data

Supplementary data (additional material related to activity calculation of penicillin derivatives (compounds **1**, **42**, **80** and **87**) and expansion of Eq. (2) for these compounds were given as an example using the κ_j values in model 1 and corresponding parameters) associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2011.02.035.

References and notes

1. Rolinson, G. N.; Sutherland, R. *Br. J. Pharmacol.* **1965**, *25*, 638.
2. Bird, A. E.; Marshall, A. C. *Biochem. Pharmacol.* **1967**, *16*, 2275.
3. Perez, M. I. B.; Rodriguez, L. C.; Cruces-Blanco, C. *J. Pharm. Biomed. Anal.* **2007**, *43*, 746.
4. Shahid, M.; Sobia, F.; Singh, A.; Malik, A.; Khan, H. M.; Jonas, D.; Hawkey, P. M. *Crit. Rev. Microbiol.* **2009**, *35*, 81.
5. Kiralj, R.; Ferreira, M. M. C. *J. Mol. Graphics Modell.* **2006**, *25*, 126.
6. Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Comput. Aided Mol. Des.* **2003**, *17*, 103.
7. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley and Sons: New York, 2000.
8. Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tine, M. R. *J. Chem. Inf. Model.* **2006**, *46*, 2030.
9. Sikora, R.; Piramuthu, S. *Eur. J. Oper. Res.* **2007**, *180*, 723.
10. Hasegawa, K.; Arakawa, M.; Funatsu, K. *Chemom. Intell. Lab. Syst.* **1999**, *47*, 33.
11. Pandey, G.; Saxena, A. K. *J. Chem. Inf. Model.* **2006**, *46*, 2579.
12. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509.
13. Duca, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367.
14. Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. *Molecules* **2010**, *15*, 3281.
15. Bersuker, I. B. *Curr. Pharm. Des.* **2003**, *9*, 1575.
16. Bersuker, I. B.; Bahceci, S.; Boggs, J. E.; Pearlman, R. S. *SAR QSAR Environ. Res.* **1998**, *10*, 157.
17. Bersuker, I. B.; Bahceci, S.; Boggs, J. E.; Pearlman, R. S. *J. Comput. Aided Mol. Des.* **1999**, *13*, 419.
18. Bersuker, I. B.; Bahceci, S.; Boggs, J. E. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1363.
19. Rosines, E.; Bersuker, I. B.; Boggs, J. E. *Quant. Struct.-Act. Relat.* **2001**, *20*, 327.
20. Makkouk, Al. H.; Bersuker, I. B.; Boggs, J. E. *Int. J. Pharm. Med.* **2004**, *18*, 81.
21. Marenich, A. V.; Yong, P. H.; Bersuker, I. B.; Boggs, J. E. *J. Chem. Inf. Model.* **2008**, *48*, 556.
22. Bersuker, I. B. *J. Comput. Aided. Mol. Des.* **2008**, *22*, 423.
23. Holland, J. H. *Adaptation in Artificial and Natural Systems*; MIT Press: Cambridge, USA, 1992.
24. Hasegawa, K.; Miyashita, Y.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 306.
25. Cho, S. J.; Hermsmeier, M. A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 927.
26. Fernandez, M.; Caballero, J.; Fernandez, L.; Sarai, A. *Mol. Divers.* **2010**. doi:10.1007/s11030-010-9234-9.
27. Sarıpınar, E.; Geçen, N.; Sahin, K.; Yanmaz, E. *Eur. J. Med. Chem.* **2010**, *45*, 4157.
28. Sahin, K.; Sarıpınar, E.; Yanmaz, E.; Geçen, N. *SAR and QSAR Environ. Res.* in press
29. Bersuker, I. B.; Dimoglo, A. S. In *The Electron-Topological Approach to the QSAR Problem, Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: USA, 1991; p 423.
30. Dimoglo, A. S.; Vlad, P. F.; Shvets, N. M.; Coltsa, M. N.; Güzel, Y.; Saracoglu, M.; Sarıpınar, E.; Patat, Ş. *New J. Chem.* **1995**, *19*, 1217.
31. Saripinar, E.; Guzel, Y.; Patat, Ş.; Yildirim, I.; Akcamur, Y.; Dimoglo, A. S. *Arzneim-Forsch/Drug Res.* **1996**, *46*, 824.
32. Güzel, Y.; Saripinar, E.; Yildirim, I. *J. Mol. Struc.-Theochem.* **1997**, *418*, 83.
33. Koçyiğit-Kaymakçıoğlu, B.; Oruç, E.; Unsalan, S.; Kandemirli, F.; Shvets, N.; Rollas, S.; Dimoglo, A. *Eur. J. Med. Chem.* **2006**, *41*, 1253.
34. <http://www.borland.com/delphi/>
35. Guner, O. F. *Pharmacophore Perception, Development, and Use in Drug Design*; International University Line: La Jolla, California, 2000.
36. Mason, J. S.; Good, A. C.; Martin, E. J. *Curr. Pharm. Des.* **2001**, *7*, 567.
37. Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. *J. Med. Chem.* **2010**, *53*, 539.
38. Altun, A.; Golcuk, K.; Kumru, M.; Jalbout, A. F. *Bioorg. Med. Chem.* **2003**, *11*, 3861.
39. Altun, A.; Kumru, M.; Dimoglo, A. *J. Mol. Struc.-Theochem.* **2001**, *535*, 235.
40. MATLAB (ver 7.0), The MathWorks Inc, 3 Apple Hill Drive, Natick, MA 01760-2098.
41. Schuurmann, G.; Ebert, R. U.; Chen, J.; Wang, B.; Kuhne, R. *J. Chem. Inf. Model.* **2008**, *48*, 2140.
42. Damme, S. V.; Bultınck, P. *J. Comput. Chem.* **2007**, *28*, 1924.
43. Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chem. Inf. Model.* **2009**, *49*, 1669.
44. SPARTAN, Version 06; Wavefunction, Inc., **2006**.
45. Zhou, Z.; Parr, R. G. *J. Am. Chem. Soc.* **1990**, *112*, 5720.
46. Hu, L.; Chen, G.; Chau, R. M. W. *J. Mol. Graphics Modell.* **2006**, *24*, 244.
47. Parr, R. G.; Szentpaly, L. V.; Liu, S. B. *J. Am. Chem. Soc.* **1999**, *121*, 1922.
48. Chattaraj, P. K.; Maiti, B. *J. Phys. Chem. A* **2001**, *105*, 169.
49. Vleeschouwer, F. D.; Speybroeck, V. V.; Waroquier, M.; Geerlings, P.; Proft, F. D. *Org. Lett.* **2007**, *9*, 2721.
50. Murray, C. W.; Verdonk, M. L. *J. Comput. Aided. Mol. Des.* **2002**, *16*, 741.
51. Fukui, K.; Yonezawa, T.; Shingu, H. *J. Chem. Phys.* **1952**, *20*, 722.
52. Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H. *J. Chem. Phys.* **1954**, *22*, 1433.
53. Woodward, R. B.; Hoffmann, R. *The Conservation of Orbital Symmetry*; Verlag Chemie: Weinheim, Germany, 1970.
54. Franke, R. *Theoretical Drug Design Methods*; Elsevier: Amsterdam, 1984.
55. Spartan 08 for Windows, Macintosh and Linux; Tutorial and User's Guide; Wavefunction, 2006.
56. Pavlov, T.; Todorov, M.; Stoyanova, G.; Schmieder, P.; Aladjov, H.; Serafimova, R.; Mekenyan, O. *J. Chem. Inf. Model.* **2007**, *47*, 851.
57. Livingstone, D. *Data Analysis for Chemists*; Oxford Univ., 1995. p 155.
58. Wold, S. *Technometrics* **1978**, *20*, 397.