Quantitative Structure−Activity Relationship of Flavonoid p56^{lck} Protein Tyrosine Kinase Inhibitors. A Neural Network Approach

M. Novič,[§] Z. Nikolovska-Coleska,[‡] and T. Solmajer*,[†]

Department of Chemometrics, National Institute of Chemistry, Ljubljana, Slovenia, Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University "St. Ciryl and Methodius", Skopje, Macedonia, and Department of Molecular Modelling and NMR Spectroscopy, National Institute of Chemistry and Lek, d.d., Ljubljana, Slovenia

Received March 7, 1997[⊗]

Specific inhibitors of protein tyrosine kinase as antiproliferative agents are instrumental in several aspects of neoplastic disease and have found wide interest as potential pharmacological agents. We have applied an artificial neural network based on a counterpropagation algorithm to develop quantitative structure-activity relationships in a large dataset of 105 flavonoid derivatives that inhibit the enzyme p56^{lck} protein tyrosine kinase. The results of such approach were compared with the linear multiregression analysis with regard to the ability to fit biological activity surfaces, predict activity, and explore the nonlinear aspects of the dependence of activity on properties. Excellent correlation was obtained for both classical and quantum chemical descriptors, and relevance of the descriptors to binding properties of the enzyme receptor active site is hypothesized.

1. INTRODUCTION

Protein-tyrosine kinases[1] (PTKs) play critical roles in both normal and neoplastic cellular signal transduction. PTKs which catalyze the transfer of the terminal phosphate of ATP to tyrosine residues on substrate proteins are key elements in these signal transduction pathways. Moreover, in many human malignancies a specific PTK is activated or overexpressed. Enhanced PTK activity resulting from tyrosine kinase overexpression can activate mutations or lead to persistent stimulation by autocrinically secreted growth factors, which in turn can lead to disease.[2] Abnormal activity of tyrosine kinases has been implicated in many cancers as well as in nonmalignant proliferative diseases such as atherosclerosis and psoriasis and in a large number of inflammatory responses.[3] The development of specific PTK inhibitors as pharmacological tools and potential antiproliferative agents has therefore become an active area of research.[4−6]

Flavonoids comprise a large group of low molecular weight substances found practically in all parts of the plants. The broad spectrum of biological activity within the group and the multiplicity of actions displayed by certain individual members make the flavonoids one of the most intriguing class of biologically active compounds, termed as "bioflavonoids".[7] A number of naturally occurring as well as synthetic flavonoids have been shown to inhibit the function of different PTKs in a manner which is competitive with respect to ATP.[8−11] These investigations provide us with important information about qualitative structure−activity relationships for flavonoids. The lack of observed well defined quantitative structure−activity (QSAR) correlations for flavonoids as PTK inhibitors[4] has prompted us to perform a thorough investigation of the available experimental database. Enzyme inhibitory activity for a series of flavonoids was measured at p56^{lck} lymphoid cell lineage-specific PTK of the *src* family which is overexpressed in several lymphomas. In our previous work[12] we have applied multiregression analysis by using classical and computed quantum chemical parameters. Our primary purpose was to elucidate the quantitative influence of the physical interactions of ligands in the p56^{lck} enzyme cavity which play the primary role in the resulting QSAR: hydrogen bonding functionality at the phenyl ring moiety and hydrophobic interactions of the chromone part of the flavonoid molecule.

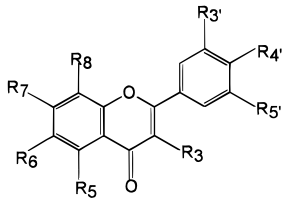

Figure 1. Molecular structure and numbering of substituents attached to the chromone moiety and phenyl ring (see also Table 1).

Artificial neural networks (ANNs) is a newly emerging field of information processing technology that has captured the interest of scientists from diverse fields.[13] Results of these studies have frequently pointed out that neural networks have been found to be extremely suitable for data processing in which causality relationships in the model environment cannot be exactly defined *a priori*. Thus its usefulness in complex biology-related responses is suggested, and as ANNs can be developed into models, they are becoming important tools in QSAR research. The neural networks are by definition computer-like processor structures derived from the simplified concept of the brain in which a number of nodes, called processing elements or neurons, are interconnected in a netlike structure. One such type of ANNs is called "error back propagation" [14] and is most widely applied in QSAR studies.[15−17] Another type of ANNs, first exploited

* Corresponding author.
[†] National Institute and Chemistry and Lek.
[‡] University "St. Ciryl and Methodius".
[§] National Institute of Chemistry.
[⊗] Abstract published in *Advance ACS Abstracts,* November 1, 1997.

S0095-2338(97)00222-9 CCC: $14.00

by Kohonen,[18] has its working principle modeled after the human learning process and the accumulation of knowledge in the brain core. These ANNs are based on unsupervised learning strategy and are suitable for initial examination of data for which the knowledge and comprehension of the responses is not needed in advance. One extension of Kohonen ANNs, found to be useful for modeling the response surfaces is counterpropagation ANN[19] (CP-ANN).

Although CP-ANNs have proved to be efficient in modeling response surfaces in various fields,[20,21] they have thus far not been extensively used in QSAR.[22] The implementation of this approach in our case study of QSAR in a large bioactive flavonoids data base confirms that such methodology could provide one with a viable and superior alternative to the error back propagation and standard Kohonen algorithms.

2. MATERIALS AND METHODS

2.1. The Data Used. A dataset[8−10] of 105 bioactive flavonoids that inhibit enzyme protein tyrosine kinase p56lck, in our view, provides a sufficiently general basis for comparison of the performance of a neural network approach against standard multiregresional QSAR. The same biological data and structural parameters that were used in the present work for a neural network approach were employed previously in a companion study of this series of compounds[12] in which the relationship of particular classical and quantum chemical descriptors in description of flavonoids inhibitory activity is discussed. The descriptors set containing both classical parameters and quantum chemical computed descriptors was chosen to describe the main physical forces that could be instrumental in inhibitor binding to the protein receptor.

Classical parameters used in this study were Hansch hydrophobic constants (π), Hammett electronic constant for the *meta* and *para* position (σ_m and σ_p), and MR (molar refractivity); all were summed for the substituents on the chromone moiety, phenyl ring, and entire molecule. These parameters were compiled by standard procedures from the literature.[23] We tested also the following computed quantum chemical parameters: net atomic charge (10 parameters) at flavonoid carbon atoms bearing a substituent, net atomic charge at substituted phenyl ring 3′ or 4′ carbon atoms $\delta_{3'4'}$, energies of the highest occupied molecular orbital (ϵ_{HOMO}), and the lowest unoccupied molecular orbital (ϵ_{LUMO}), surface density, surface *HOMO* density, surface *LUMO* density, total dipole moment (μ), and the energy barrier of the rotation of the phenyl ring (ϵ_{rot}) about the C2−C1′ bond. In order to compute the quantum chemical descriptors itemized above we have used the well documented semiempirical AM1 approach.[24,25] The orientation of the phenyl-ring with respect to the chromone moiety of the flavonoid structure requested prior determination of the minimum energy conformation for rotations about the C2−C1′ bond. Thus, all the quantum chemical descriptors were calculated with AM1 method using the fully optimized geometry of the compound in question.

A novel descriptor $\delta_{3',4'}$ introduced above deserves some additional explanation. Electrostatic potential surfaces have proven to be a valuable measure of interaction patterns of ligands at a macromolecular receptor active sites which contain a hydrogen bond donor−acceptor functionality.[26,27] By superimposition of the electrostatic potential surface encoded onto electron density surface of the substituted

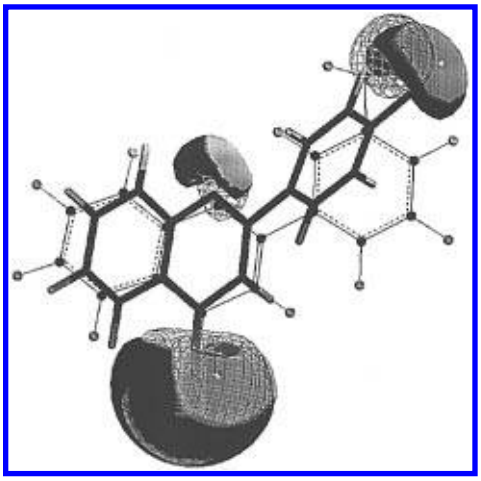

Figure 2. Superimposition of 3′ substituted (thin lines) and 4′ substituted flavonoids (full lines) by overlapping of the molecular electrostatic potential encoded onto total electron density surfaces. 3′ OH flavone MEP is represented by mesh, 4′-OH by transparent surface. Isopotential value of −20 kJ/mol is plotted in both cases.

phenyl ring around e.g., 3′-OH and 4′-OH substituents, respectively, it is shown that a very similar interaction pattern is obtained from either 3′-OH or 4′-OH substituent in the whole flavonoid structure (at chromone oxygens at position 1 and 4 as well). Also, based on qualitative SAR it is reasonable to expect that in a limited space which would be occupied by a substituent of 3′ and 4′ carbon atoms of the phenyl ring only one hydrogen bond is available for simultaneous interaction at the putative enzymatic receptor site. In consequence by using the atomic charge either at substituted C(3′) or C(4′) atom of the bulk of the electrostatic potential value in this section of space around phenyl ring moiety is accounted for, and a single descriptor $\delta_{3'4'}$ provides a reasonable measure for electrostatic interaction in this region.

2.2. Counterpropagation Artificial Neural Networks. The architecture (layout) of the CP-ANNs is composed of two layers: the input (Kohonen) and the output (Grossberg) layer. The neurons in both layers are arranged in a ($N_{net} \times N_{net}$) dimensional map. Each layer is placed exactly one above the other, having the same number of neurons (Figure 3). The CP-ANN is capable of solving the supervised type of problems like modeling the response surfaces for the set of *n*-dimensional objects X_s (x_{s1}, $x_{s2}...x_{si}...x_{sn}$) with known *m*-dimensional responses Y_s (y_{s1},$y_{s2}...y_{si}...y_{sm}$). The molecular structures X_s of the series were initially represented in two ways: by 11 classical ($n = 11$) and by 18 quantum chemical descriptors ($n = 18$). Each neuron in the input layer W_j^{inp} (w_{j1}, $w_{j2}...w_{ji}...w_{jn}$) has *n* weights since it must be comparable with the *n*-dimensional object representation. The neurons in the output layer are *m*-dimensional W_j^{out} (w_{j1}, $w_{j2}...w_{ji}...w_{jm}$) like response vectors. The distribution of the *m* weights in the ($N_{net} \times N_{net}$) output layer constructs *m* response surfaces, one for each of the *m* sought properties. The response (Y_s) for supervised training of the CP-ANN is inhibitory activity against protein-tyrosine kinase p56lck of each investigated compound. The network is trained in the training procedure in which the weights are corrected for each new input object. The amount of correction is scaled with the two correction factors, a_{max} and a_{min}. The input to

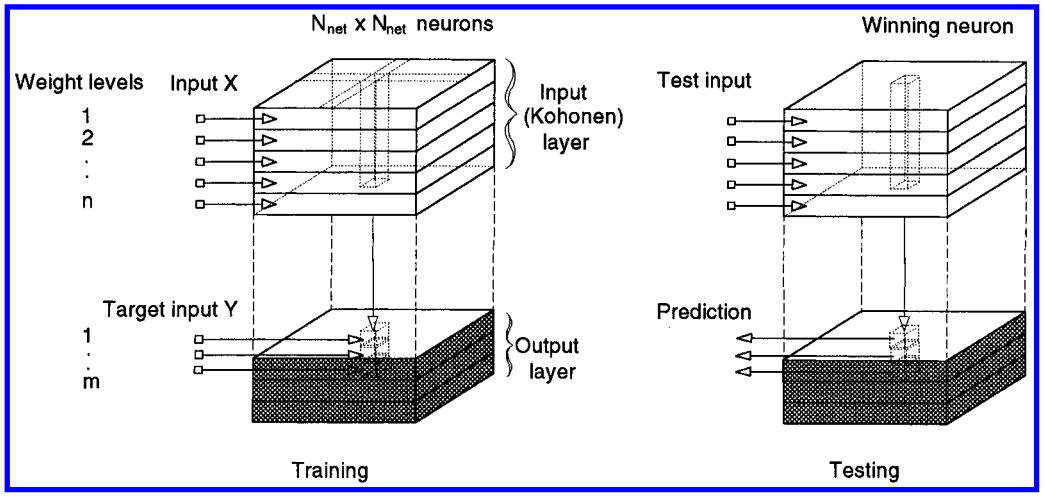

Figure 3. The CP-ANN shown schematically for training and testing procedure. The neurons are arranged in $N_{net} \times N_{net}$ blocks, divided into two layers: input (Kohonen) and output layer.

the entire set of training objects and adequate correction of weights is called *one epoch* of training process.[13]

Two important issues need to be addressed in each application of ANNs: careful optimization of the appropriate training time and proper selection of the training set. The CP-ANNs usually require only several hundred epochs to be properly trained and one of the main advantages offered by the CP-ANNs is an order of magnitude lower number of epochs compared with learning times of the more frequently used error-back-propagation algorithm. However, the increasing number of training epochs (training time) does not improve the prediction results of the test-set although the prediction results of the training set can further be improved. This is a well-known handicap of overtrained networks. For example, when 200 epochs are increased up to 1000, the quality of retrieval, i.e., prediction of the properties of the training set, is raised but the performance of the test set fit is increasingly poor. Since the performance of the final CP-ANN model is evaluated by the statistical parameters obtained from the test set objects, the optimal number of epochs cannot be determined on the basis of minimal threshold training error−RMS of the training set alone. While this RMS value should be kept reasonably low, a compromise in choosing the statistical description which fits both training and test data set needs to be found.

Selection of data for the training set from the data base of all compounds is crucial for the evaluation of the model. One would like to ensure that the training set is represented in the model in such a way that the predicted properties of the compounds in the test set are within statistical limits determined by such procedure. In our case it was done by a method recently developed in this laboratory and based on Kohonen mapping.[20] In this selection procedure all compounds from the data set, represented by classical or quantum-mechanical parameters, were initially used for training of the Kohonen ANN. Thus they were located in the area of $N_{net} \times N_{net}$ neurons in the same way as they were placed in the input layer of the CP-ANN, according to the "best match" criterion. Objects with similar representation by a chosen parameter set are located close together. If the neurons in the $N_{net} \times N_{net}$ area are labeled by tags carrying the number of objects exciting them, a labeled map is generated called *top-map*. In order to obtain the most representative selection of the training set one has to pick those objects that are evenly distributed throughout the whole *top-map*.

The proposed architecture of the CP-ANN was (10 × 10) neurons with 11 weights (classical descriptors) in the input layer and one weight in the output layer [10 × 10 × (11 + 1)]. The CP-ANNs were trained for 200 epochs, parameters a_{max} and a_{min} being 0.5 and 0.01, respectively. In the output layers of all constructed models the weights were adapted to biological activity representing the modeled response surface.

3. RESULTS AND DISCUSSION

In Tables 1a−c a summary of the results for various models is given along with the structure of molecules that are components of the enzyme inhibitors database consisting of 105 inhibitors. A large proportion of the compounds (49 in total have biological activity of 2.70) are on the lower bound for the assay used. These compounds may not be well discriminated in terms of their physicochemical differences. Thus, to ascertain this issue we have deleted the majority (35 - tagged with an asterisk in Table 1) of inactive compounds from the full data set and compiled a subset with 70 compounds (54 active and 16 inactive). These results are presented in Section 3.3.

In order to further locate the source of deviation from ideal correlation and connect it to the chemical structure of compounds we further dissected the total data set in three subsets with related structures (Tables 1a−c), according to the work of Cushman at al.[8−10] It was anticipated that within the subsets better correlation between structural representation and biological activity could be found because of similar mechanisms responsible for the activity of compounds in the study. We discuss the development of the classical and quantum chemical models and resulting correlations between biological activity and calculated models in three sections: in Sections 3.1 and 3.2, an evaluation of the classical and quantum chemical models, respectively, is given for the complete data set of 105 flavonoids, and in Section 3.3 the subset models are discussed separately. With the use of *top-map* approach (Figure 4) we have separated the molecules in the data set into two lists: 54 compounds distributed to cover the whole 10 × 10 map were selected for the training set and the remaining 51 compounds served for testing the model obtained by CP-ANN. Also, in order to avoid the bias of epoch number determination to the actual test set of compounds a new test set (control set) was selected from the objects in the original training set. The training of

Table 1. Biological Activity [log ($1/IC_{50}$)] − Comparison of Experimental Values for the Entire Set of Compounds and Predicted Values Obtained by the Models CP-M3 and CP-M3-qm for the Entire Set and by (a) the model CP-M-S1-3 for the First Subset, (b) the Model CP-M-S2-6 for the Second Subsets, and (c) the Model CP-M-53-6 for the Third Subset

(a)

			CP-M3		CP-M3-qm		CP-M-S1-3	
no.	substitutents	exp[a] log($1/IC_{50}$)	training	testing	training	testing	training	testing
1	3,5,7,3′,4′-OH	4.88	4.85		4.84		4.88	
2	3,7,3′,4′-OH	4.86		5.11	4.82		4.86	
3	5,7,4′-OH	4.83	4.83		4.79		4.83	
4	5,4′-OH	4.80	4.17		4.77		4.80	
5	6,3′-OH	4.80		4.17	4.36			3.89
6	5,7-OH	4.71	4.67		4.62		4.71	
7	5,7,3′,4′-OH	4.46		5.11	4.44			4.51
8	7,3′-OH	4.41		4.67		4.36		4.36
9	6-OH,3′,4′,5′-OCH3	4.22	4.19		4.20		3.46	
10	3,5,7,4′-OH,3′,5′-OHC3	4.16	4.14			4.84	4.16	
11	3,5,7,3′,5′-OH	4.00		4.85	4.02			4.88
12	6,4′-OH	3.93	3.93		4.36			3.89
13	7,8,4′-OH,3′,5′-OCH3	3.92	3.92		3.66		3.92	
14	6-OH,4′-OR	3.92	3.67		3.90		3.92	
15	6,4′-OH,3′,5′-OCH3	3.89	3.89		3.90		3.89	
16	7,4′-OH	3.78		4.17		4.36		4.80
17	7,8,3′OH	3.75		4.83	3.75			4.37
18	3,5,7-OH	3.53		3.85	3.57			4.31
19	5,4′-OH,7-OCH3	3.55	4.17			3.86		4.80
20	5,3′-OH	3.50		4.67		4.77		4.36
21	7,8-OH	3.50		4.17	3.52			4.05
22	7-OH	3.47	3.41		3.08		3.47	
23	6-OH,3′,5′-OCH3,4′-OR	3.43	3.67		3.41		3.43	
24	7,8-OH,3′,4′,5′-OCH3	3.40	3.39		3.66		3.40	
25	7-OH,4′-OR	3.01		3.67		3.90		2.82
26	7,4′-OH,3′,5′-OCH3	2.90	3.39			3.90		4.80
27	7-OH,3′,5′-OCH3,4′-OR	2.82		3.67		3.41	2.76	
*28	7-OH,4′-OBn	2.69	2.69		2.69		2.69	
*29	7,8,3′,4′,5′-OCH3	2.70	2.70		2.70		3.46	
*30	7,8-OH,3′,5′-OCH3,4′-OR	2.70	2.72		2.70		2.70	
*31	7,8-OAc, 3′,5′-OMe, 4′-OR	2.70		2.72		2.70	2.70	
*32	6,3′,4′,5′-OCH3	2.70		2.70	2.70			2.70
*33	7-OH,3′,4′,5′-OCH3	2.70		2.99		2.77		3.84
*34	7-OAc,3′,5′-OCH3,4′-OH	2.70		3.07		2.70		4.80
*35	7-OAc,3′,5′-OCH3,4′-OR	2.70		3.67	2.71		2.70	
*36	7,3′,4′,5′-OCH3	2.70		2.70	2.74			2.70
*37	5-OH,4′-OBn	2.70		2.69		2.69		2.69

(b)

			CP-M3		CP-M3-qm		CP-M-S2-6	
no.	substituents	expl[a] log(I/IC_{50})	training	testing	training	testing	training	testing
*38	6-OH,4′-NH2	5.92	5.92		5.82		5.92	
39	5,7-OH,4′-NH2	5.13	5.11		5.05			5.92
40	4′-OH,3′,5′-OCH3	4.57	4.51		4.53		3.93	
41	7-OH,4′-NH2	3.86		5.43		4.01		5.92
42	4′-NH2	3.68		3.93	3.69		3.68	
43	3-COOMe,4′-OH	3.36	3.41		3.36		3.36	
44	4′-OH	3.30		3.41	3.33		3.93	
45	3-COOMe,4′-NH2	3.09		4.17	3.15			3.08
46	3-COOH, 7-OMe,4′-OH	2.99	3.01		3.00			3.16
47	3-COOH,4′-OH	2.80		3.01		3.28	2.80	
*48	3′,4′,5′-OCH3	2.70	2.70		2.73			3.35
49	3-COOMe,3′,4′,5′-OMe	2.70	2.70		2.70		2.70	
50	3-COOMe,3′,5′-OMe	2.70	2.70		2.70		2.70	
51	3-COOMe,3′,4′-OMe	2.70		2.70	2.70			2.70
*52	3-COOMe,4′-OMe	2.70	2.72			2.70		2.70
*53	3-COOMe,4′-Br	2.70	2.70		2.70		2.70	
54	3-COOMe,4′-OBn	2.70	2.71		2.70		2.70	
55	3-COOMe,7-OMe, 4′-OBn	2.70		2.71	2.70			2.70
56	3-COOMe,6-OMe, 4′-OBn	2.70		2.71		2.70		2.70
*57	3-COOMe,4′-NO2	2.70	2.70			2.70	2.70	
*58	3-COOMe,7-OMe,4′-NO2	2.70		2.70	2.70			2.70
*59	3-COOMe,6-OMe, 4′-NO2	2.70		2.70		2.70		2.70
*60	3-COOMe,5,7-OBn,4′-NO2	2.70	2.72		2.70		2.70	
*61	3-COOH,3′,4′,5′-OMe	2.70	2.70		2.70		2.70	
*62	3-COOH,3′,5′-OMe	2.70	2.70		2.70			2.70
*63	3-COOH,3′,4′-OMe	2.70		3.05		2.70		2.70
*64	3-COOH,4′-OMe	2.70		2.72		2.70	2.70	
65	3-COOH,4′-Br	2.70		2.70		2.70	2.70	
66	3-COOH,4′-NO2	2.70	2.70			2.82	2.70	

Table 1. (Continued)

(b) (continued)

			CP-M3		CP-M3-qm		CP-M-S2-6	
no.	substitutents	exp[a] log(l/IC_{50})	training	testing	training	testing	training	testing
67	3-COOH,7-OMe, 4′-NO2	2.70	2.70		2.70			2.70
68	3-COOH,6-OMe, 4′-NO2	2.70		2.70		2.70		2.70
*69	3-COOMe,7-OMe, 4′-OH	2.70		3.41		2.70	2.70	
*70	3-COOMe,6-OMe, 4′-OH	2.70		3.41		2.70		2.70
*71	3-COOMe,7-OMe, 4′-NHAc	2.70	2.71		2.70		2.70	
*72	3-COOMe,6-OMe,4′-NHAc	2.70		2.71	2.70			2.70
73	3-COOH,5,7-OH,4′-NO2	2.70	2.70		2.70		2.70	
74	4′-NO2	2.70	2.70		3.08			2.70
*75	7-OH,4′-NO2	2.70	2.70		2.73			2.70
*76	6-OH,4′-NO2	2.70		2.70	2.72			2.70
*77	5,7-OH,4′-NO2	2.70	2.72			3.52		2.70

(c)

			CP-M3		CP-M3-qm		CP-M-S3−6	
no.	substituents	expl[a] log(I/IC_{50})	training	testing	training	testing	training	testing
78	6-OH, 5, 7, 4′-NH2	4.74	4.73		4.11		4.74	
79	6-OH, 5, 7, 3′-NH2	4.34		4.73	4.33			4.74
80	6-OMe, 8, 3′-NH2	4.25	4.24		4.23		4.25	
81	6,4′-NH2	3.99	3.99		4.01		3.99	
82	6, 8, 4′-NH2	3.97	3.99		3.69		3.97	
83	6-OH, 8, 4′-NH2	3.93	3.89			5.05	3.93	
84	8,4′-NH2	3.91		3.99	3.91			3.99
85	7-OH, 6, 4′-NH2	3.85	3.89			5.05	3.85	
86	6,3′-NH2	3.70	3.72		3.75			3.99
87	5-OH, 6, 4′-NH2	3.65		3.89		5.05		3.85
88	5-OH, 8, 4′-NH2	3.49		3.89	4.11			3.85
89	7-OH, 8, 4′-NH2	3.48		3.89		3.69		3.85
90	6-OMe, 8, 4′-NH2	3.42		3.99	3.69		3.42	
91	7-OH, 6, 3′-NH2	3.30	3.32		3.35		3.30	
92	7-OH, 6, 8, 4′-NH2	3.12		4.73		4.33	3.12	
93	7-OH, 6, 8, 4′-NO2	2.81	2.81		2.81		2.81	
94	5-OMe, 8, 4′-NH2	2.79		4.24	2.82			3.42
95	7-OH,8,4′-NO2	2.73	2.73		2.72		2.73	
96	6,4′-NO2	2.70	2.70			2.72	2.70	
97	8,4′-NO2	2.70		2.70		3.08		2.70
*98	7-OH, 6, 4′-NO2	2.70		2.73	2.72			2.73
*99	5-OH, 8, 4′-NO2	2.70		2.73		2.73		2.73
100	6-OMe, 8,4′-NO2	2.70	2.70			2.70	2.70	
101	5-OMe, 8, 4′-NO2	2.70		2.70		2.72		2.70
*102	6,8,4′-NO2	2.70	2.70			2.70	2.70	
*103	6-OH, 8, 4′-NO2	2.70	2.70			2.72	2.70	
*104	5-OH, 6, 4′-NO2	2.70		2.73	2.73			2.73
*105	6-OH, 5, 7, 4′-NO2	2.70		2.81		2.70		2.81

[a] IC_{50} is the molar concentration of the flavonoids necessary to give half-maximal inhibition as compared to control assay carried out in the absence of inhibitor, but in the presence of DMSO carrier. The biological activity values were taken from the works by Cushman and co-workers.[8−10] [b] Compounds labeled by an asterisk were excluded in the reduced set of 70 compounds used for validation procedure.

0	0	0	0	0	0	0	0	0	0
2				0					1
0		0	0		0		0	0	0
				5			0		0
0		0	3	1		0	0		0
		3			0		0	0	0
	3			4			0		1
3		2		4		3		2	
		2			3			3	3
2		3		3	1		3		3

Figure 4. Top-map obtained from (10 × 10) Kohonen ANN trained with 105 flavonoid derivatives represented with classical parameters. The neurons tagged with labels (numbers from 0 to 5 indicating six activity levels) were excited by at least one compound (see Table 1a−c, column CP-M3 denoted by training). Fifty-four neurons in total were occupied.

the optimal number of epochs was done on a subset of 66 objects selected out of the full set of 105 objects.

Table 2. The Dependence of Resulting Correlation[a] between the Experimental and Predicted Biological Activities from CP-ANN Model[b] with Quantum Chemical Parameters on the Number of Epochs Used for Training

epochs	$r^2_{tr\text{-}33}$	$r^2_{control}$	$SD_{control}$	$f_{control}$
100	0.9819	0.6618	0.45	60.7
150	0.9943	0.6550	0.50	58.9
200	0.9982	0.7060	0.45	74.4
250	0.9995	0.6638	0.48	61.2
300	0.9998	0.6262	0.53	51.9
600	1.0000	0.6620	0.51	60.7
1000	1.0000	0.6171	0.52	50.0

[a] The correlation is expressed with statistical parameters: correlation coefficient r^2, standard deviation SD, and factor f. [b] Three quantum-mechanical parameters were used for structure representation. The new training set of 33 object out of original 66 training objects was selected, leaving the rest of the 33 objects for the control set.

The results in Table 2 are shown with different number of epochs used for training. To ensure that the ANN was not overtrained a variety of epochs (100−1000 epochs) was

Table 3. The Resulting Correlation[a] between the Experimental and Predicted Biological Activities for 105 Compounds from CP-ANN Models from Classical and Quantum Chemical Parameter Set[b]

model	parameters	$r^2_{training}$	r^2_{test}	SD_{test}	f_{test}
CP-M11	π_{ch}, π_{ph}, $\Sigma\pi$, $(\Sigma\pi)^2$, σ_{ch}, σ_{ph}, $\Sigma\sigma$ MR_{ch}, MR_3, MR_{ph}, ΣMR	0.97	0.69	0.49	110.0
CP-M6	π_{ch}, π_{ph}, $(\Sigma\pi)^2$, σ_{ch}, σ_{ph}, MR_3	0.92	0.56	0.55	62.2
CP-M5	$\Sigma\pi$, $(\Sigma\pi)^2$, σ_{ph}, MR_3 ΣMR	0.97	0.67	0.51	100.5
CP-M3	$(\Sigma\pi)^2$, $\Sigma\sigma$, MR_{ph}	0.96	0.70	0.47	112.1
CP-M3-qm	SA, Σ_{3-8}, $\delta_{3',4'}$	0.96	0.78	0.40	127.8

[a] The correlation is expressed with statistical parameters: correlation coefficient r^2, standard deviation SD, and factor *f*. [b] The representation of compounds with 11, 6, 5, and 3 classical and three quantum chemical parameters, respectively.

attempted for the evaluation of the model with quantum chemical input neurons CP-M-qm3. Dependence of correlation coefficient of the test set on the variation of the number of epochs shows that by increasing the training time a better fit in the training set is achieved. However, in the control set this improvement is not matched by appropriate increase of the correlation coefficient. Rather, a declining $r^2_{contr} = 0.62$ in the case of $N = 300$ when compared with $r^2_{contr} = 0.71$ at $N = 200$ shows a familiar feature of overtrained ANN. Thus a compromise was made based on the result of this computational experiment. In all subsequent ANN trainings we have employed $N = 200$ which has shown a slightly better standard deviation of the test set $SD_{contr} = 0.45$ in comparison with $SD_{contr} = 0.48$ for $N = 250$. In summary, the correlation coefficient $r^2_{contr} = 0.71$ for the CP-M-qm3 provides a reasonable predictive power of the inhibitory activity profile of the—to our best knowledge—largest series of flavonoids considered so far.

3.1. Classical Parameter Set. The resulting CP-ANN model (CP-M11) which used the full number of classical parameters (11 weights of the input neurons) was examined first with the compounds from the training set in an attempt to retrieve the biological activities of "known" compounds. The retrieval results reflect the successfulness of training but not the quality of the model. To test the quality of the obtained model, we have to examine the predictions of activity of "unknown" compounds that were not included in the training set. For this purpose 51 compounds from the test set were input to the CP-ANN model, and the prediction of 51 biological activities were compared to the experimental values as listed in Table 1a−c. The goodness of fit between experimental and calculated bioactivities is given in terms of the squared correlation coefficients. In order to improve the starting CP-ANN model, optimization by variation of the number of input parameters was performed. Three additional CP-ANN models were built with the same number of neurons (10×10) but a different number of weights in the input neurons, corresponding to reduce number of descriptors. The results of four different models trained with the same compounds but represented by 11, 6, 5, and 3 input parameters, i.e., descriptors, are shown in Table 3.

The reduced number of parameters was determined by inspection of all 11 weight maps of the input layer and of the response surface from the output layer. We focused our attention to two factors: correlations of the weight maps of input parameters, which make some of the parameters redundant (correlated with each other), and correlations of the weight maps of input parameters with the response surface (partially overlapping contours of local maxima and minima of the weight map and the response surface) which favor a particular parameter. In the search for correlation with the response surface the best coefficients were found for the contour maps of the following input parameters, $\Sigma\sigma$-($r = 0.65$), σ_{ph}($r = 0.65$), and σ_{ch}($r = 0.57$). On the other hand, the following pairs of input parameters were found to be correlated: π_{ch} with $\Sigma\pi$ ($r = 0.84$), π_{ph} with MR_{ph}($r = 0.83$), $\Sigma\pi$ with ΣMR($r = 0.73$), π_{ph} with $\Sigma\pi$($r = 0.83$), and σ_{ch} with $\Sigma\sigma$($r = 0.94$) and thus were found to be redundant.

For the first reduced model CP-M6, the six parameters (Table 3, second row) were chosen on the basis of the corresponding six weight maps which were least correlated among themselves. The second reduced model CP-M5 with five-parameters reduced representation was computed in order to allow comparison with the best model determined in previous work with the same data but using classical multiple regression (MLR) method[12] and provides additional comparison with the results obtained from the CP-ANN modeling procedure. It is gratifying to observe that the squared correlation coefficient $r_{test}^2 = 0.67$ from CP-ANN model compares favorably with the $r^2 = 0.51$ obtained in the MLR procedure.

In order to additionally reduce the number of variables and find the most important parameters of structure representation which show correlation with biological activity we built the third optimized model (CP-M3) with the following three parameters: π^2, $\Sigma\sigma$, and MR_{ph}. The parameter $\Sigma\sigma$ reveals best single correlation of corresponding weight map with the response-map as can also be seen from its correlation coefficient $r = 0.65$. The analogous weight maps for the parameters π^2 and MR_{ph} differ most from the weight maps of other parameters what indicates the additional content of information for these structural features.

The synergistic effect of a second input parameter can be explored by considering the response surfaces with respect to two input parameters, while the remaining *n−2* parameters are kept constant. Three such graphs for the CP-M5 model are presented in Figure 5. In Figure 5a the combined effect of parameters $\Sigma\pi$ and σ_{ph} on the response surface is plotted. It can be stated that the plateau of response surface with the maximal values is in the region of the parameters $\Sigma\pi$ below -1.0 and σ_{ph} below 0. Similarly, in Figure 5b we compare the influence of parameters $\Sigma\pi$ and ΣMR. The optimal region of the parameter $\Sigma\pi$ is the same as previously (Figure 5a), whereas the values of parameter ΣMR must be rather low, below 2.0. This is also in accordance with the third graph where the dependence of the biological activity is plotted against the parameters ΣMR and σ_{ph} (Figure 5c). The optimal regions of the parameters ΣMR and σ_{ph} are below 2.0 and below 0.3, respectively. Thus, by such simple graphical procedure one can estimate the optimal values of parameters resulting from CP-ANN process.

3.2. Quantum Chemical Parameter Set. We repeated all the above procedures with a set of 18 quantum-mechanical parameters as input variables. The reduction of all 18 input parameters representing input neurons described in the Methods section was made in the first step. Only three parameters, found to contain the most information and to be noncorrelated, were chosen: surface area SA, sum of charges in the chromone moiety Σ_{3-8}, and net atomic charges of carbon atoms in phenyl ring $\delta_{3'4'}$ (model CP-M3-qm). With these parameters the procedure for the training set selection based on Kohonen ANN was repeated. This resulted in a top-map similar to the one obtained with classical parameters (see Figure 4). The classical and quantum chemical top-

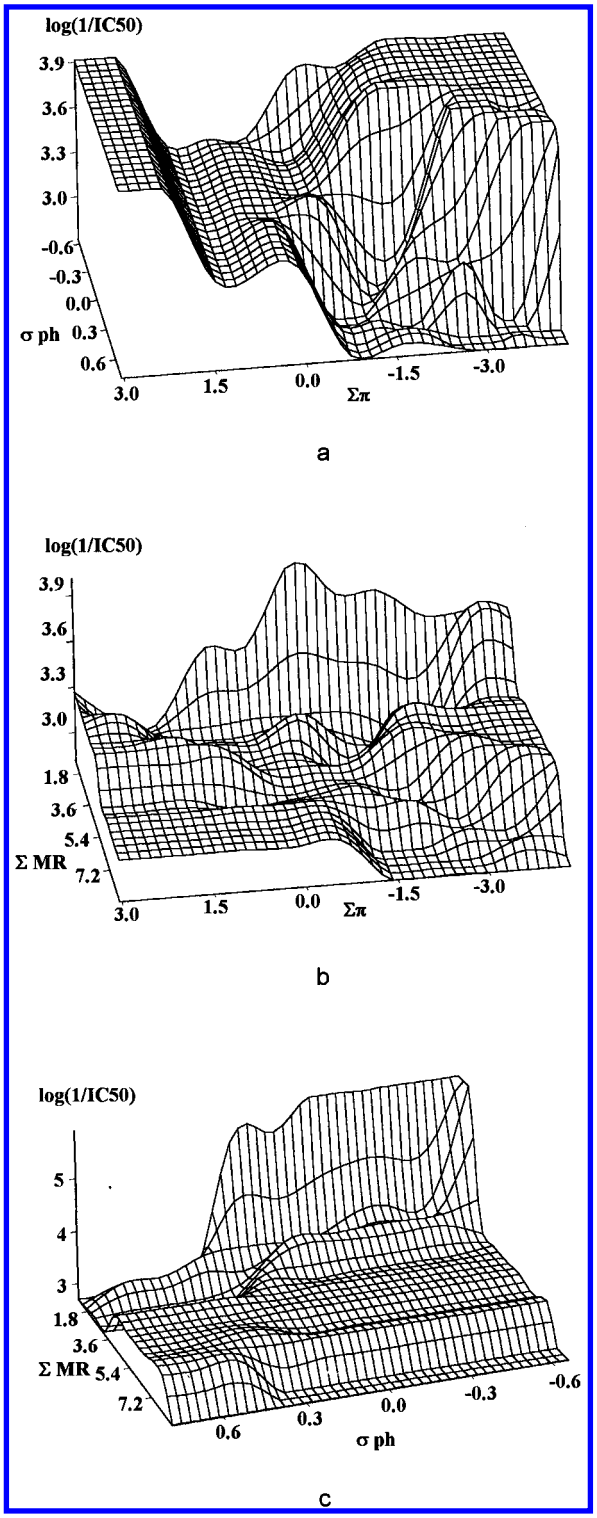

Figure 5. Response surfaces as a result of CP-ANN model when using classical parameters for the structure representation. Each plot illustrates the dependence of the modeled activity on the two classical parameters (the two axes labels), while the other parameters are kept constant at the median value.

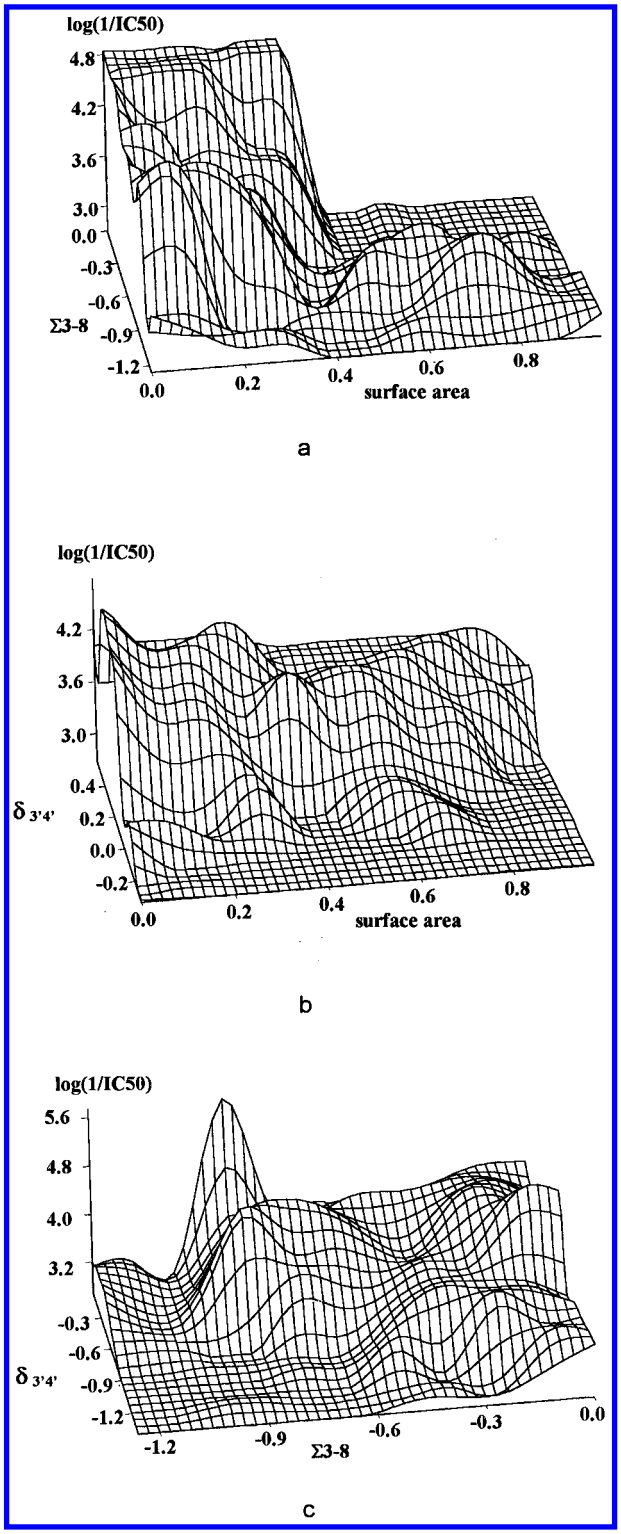

Figure 6. Response surfaces as a result of CP-ANN model when using three quantum-mechanical parameters for the structure representation. The three plots illustrate the dependence of the modeled activity on the two quantum-mechanical parameters, denoting the two axes, the third parameter being constant.

maps differ by the number of neurons excited at least once. In the latter case 66 neurons were excited at least once yielding 66 compounds for the training set. This training set was further divided into a true training set and control of 33 compounds each. The remaining 39 compounds were employed in the test set. The statistical evaluation of predicted bioactivities of the compounds in the training and test sets are shown in Table 3, bottom row.

In Figure 6 the optimal values of a single quantum chemical descriptor were determined. Here variations of response surface, as predicted by the CP-ANN model, caused by variations of the two chosen quantum-mechanical parameters, the third being constant, is presented. The three graphs are compared, and some conclusions described below can be drawn from them. The response surface in Figure 6a shows maximal values in the upper-left corner, where the values for the surface area SA are below 0.4, and for the $\sum_{3-8}$ above −0.3. The graph in Figure 6b which presents the combination of the surface area and $\delta_{3'4'}$ parameters,

respectively, shows that the best response for the parameter $\delta_{3'4'}$ is obtained for values between 0.1 and 0.3, while the surface area has almost no influence. We have arbitrarily chosen a constant value for the third parameter, $\sum_{3-8}$, to be −0.71 for this graph (in accord with the conclusions from Figure 6a this value should be less than −0.3). The third graph exhibits the largest activity (above 5.6) in a narrow region with parameter $\sum_{3-8}$ close to −0.9 and $\delta_{3'4'}$ larger than 0.4.

Finally, in order to validate the maximal expected error of the CP-ANN model, the "leave one out" procedure was performed.[13,28] One hundred five models were built, each with a training set of 104 compounds, while one compound (a different one in every case) served for the test. The cross-validation estimate for prediction error was determined by eq 1

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{-n(i)})^2 \qquad (1)$$

where n is equal to the number of compounds in the data set ($n = 105$), y_i is the activity experimental value of the ith compound, while $y_i^{-n(i)}$ is predicted activity obtained from the model which had been built with the data set not containing the ith compound. An estimate for prediction error CV = 0.34 in logarithmic units for predicted biological activity IC_{50} was obtained thus providing a statistical estimate of the predicting performance of the method used. For the reduced set of 70 compounds the cross-validation estimate for prediction error was found to be similar CV = 0.33.

The predictive residual sum of squares (PRESS), standard error of predictions (SDEP), and squared correlation coefficient of the predictions (Q^2) were 12.39, 0.56, and 0.79, respectively. However, it has to be stressed here, that the validation of models described above (Table 3) was done with the test set compounds, after approximately one half of the compounds in the data set have been carefully chosen and put in the training set as described above in the Methods Section (Figure 3) and elsewhere.[20] The methodology of automatic choice of compounds for the training set is an integral part of modeling procedure, and the CV statistical evaluation should be done, in order to be comparable with the statistical parameters of constructed models shown in Table 3, only for the test compounds.

3.3. Subset Evaluation. The compounds, listed in Table 1, are ordered with respect to the three subsets matching the compounds lists from the literature.[8−10] In general, these compounds are more similar inside each subset, and hence each subset provides a better insight into structural features relevant for binding to PTK enzymes

1. subset (S1, 37 compounds)—no. 1−37
2. subset (S2, 40 compounds)—no. 38−78
3. subset (S3, 28 compounds)—no. 79−105

In addition we have made a test with a subset of 70 compounds (not tagged with an asterisk in Table 1a−c) to determine the possibility of bias toward inactive compounds in the full data set of 105 compounds. CP-ANN model with three classical descriptors was used: the training set consisted of 40 compounds chosen according to the Kohonen mapping procedure and the remaining 30 were in the test set. The resulting statistical data, correlation coefficient $r = 0.82$, value of F-test $F = 55.8$, and standard deviation SD = 0.47, are similar to the corresponding values of the full set $r = 0.84$, $F = 112.1$, and SD = 0.47 and provide additional validation of the use of CP-ANN modeling procedure.

Table 4. Comparison of the Resulting Correlation between the Experimental and Predicted Biological Activities from CP-ANN Models for Three Subsets (S1 = 37 Compounds, S2 = 40 Compounds, and S3 = 28 Compounds) of Flavonoid Derivatives[a] for Classical and Quantum Chemical Computed Parameters

model	$r^2_{training}$	r^2_{test}	SD_{test}	f_{test}
CP-M-S1-11[b]	0.97	0.65	0.42	29.1
CP-M-S1-6[b]	0.85	0.21	0.72	4.2
CP-M-S1-3[b]	0.91	0.65	0.42	29.1
CP-M-S1-qm3[c]	0.92	0.49	0.40	15.6
CP-M-S2-11[b]	1.00	0.84	0.40	98.7
CP-M-S2-6[b]	0.94	0.85	0.38	110.7
CP-M-S2-3[b]	0.98	0.52	0.56	20.6
CP-M-S2-qm3[c]	0.80	0.56	0.50	24.0
CP-M-S3-11[b]	0.95	0.84	0.29	58.1
CP-M-S3-6[b]	1.00	0.93	0.19	153.1
CP-M-S3-3[b]	1.00	0.80	0.34	43.3
CP-M-S3-qm3[c]	0.95	0.91	0.19	111.9

[a] The correlation is expressed with statistical parameters: correlation coefficient r^2, standard deviation SD, and factor f. [b] The representation of compounds with 11, 6, and 3 classical parameters. [c] The representation of compounds with three quantum mechanical parameters.

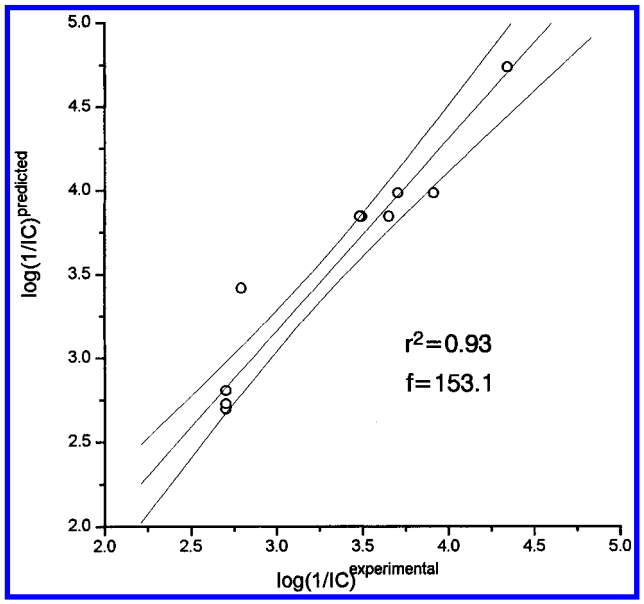

Figure 7. Correlation between experimental and predicted values of −log IC_{50} for the test set (13 compounds) of Subset 3 (model CP-ANN-qm3, see Table 4 and Section 3.3).

The CP-ANN models for subsets 1−3 were generated with the same procedures as described for the complete data set, from selection of training sets for each subset, to optimization of input parameters. The number of input neurons was smaller, only (6 × 6) because of smaller data sets. Other ANN parameters were identical to those used in previous models. The results are compiled in Table 4. Inspection of Table 4 shows that the best correlation with $r_{test}^2 = 0.93$ for six classical parameters (model CP-M-S3-6) and $r_{test}^2 = 0.91$ for three quantum chemical parameters set (model CP-M-S3-qm3) is obtained for subset S3 comprising a well balanced set of amino (active) and nitro (inactive) substituted flavonoids. In Figure 7 the experimental vs computed bioactivities are given for subset S3 and model CP-M-S3-qm3. Also in this group of compounds practically no bulky substituent groups were present, which particularly at position

3 of the chromone moiety complicate the QSAR in subset S1 for which the lowest correlation data were obtained. Although our parameter sets appear to provide for the effect of substituent size, all features introduced into binding of such flexible moiety to the receptor cavity cannot be completely adequately described. In particular, entropic effects introduced by flexible bond rotors of the substituents are difficult to estimate within a reasonably inexpensive computational procedure which are only applicable to a larger series of molecules.

Structural determinants which were found to be significant by the ANN technique can be discussed in connection with a hypothesized three-dimensional fit of these molecules to the active site of the enzyme. Both classical and quantum parameters match nicely in terms of underlying physical forces for the binding process. Surface area of the molecule is the molecular descriptor calculated by quantum chemical approach which is closely related to the classical parameters π and MR. Sum of charges on carbon atoms of the chromone moiety represents the equivalent of the classical Hammett constant σ. In addition to this, we introduced a quantum chemical descriptor $\delta_{3'4'}$, which implements the ability of the inhibitor to form hydrogen bonding at a specific spatial location, i.e., at either position 3′ or 4′ of the phenyl ring. Both options were introduced in order to simultaneously accommodate the ability of flavonoids to exhibit superposition of critical molecular functionalities in multiple orientations.

Such limited set of descriptors provides a reasonable QSAR model of a large flavonoids data set. Moreover, such model can be interpreted in accord with the qualitative proposal of Cushman et al.[8] based on inhibitory activities of flavonoid derivatives with hydroxyl and methoxy substituents at chromone and phenyl rings, respectively. Their conclusion was that at least two hydroxyl groups are necessary, positioned at both chromone and phenyl rings. Our parameter $\delta_{3'4'}$ provides an extension of this rule to include a much larger set of any substituent capable of hydrogen bond donor−acceptor functionality and which does not exceed the upper limit for the substituent size as given with the value MR = 2.0.

4. CONCLUSIONS

We have used a CP-ANN technique to study the correlation of inhibitory activity of data set of 105 flavonoid inhibitors at the enzyme $p56^{lck}$ protein tyrosine kinase. Both classical and quantum chemical descriptors were studied. The best overall correlation is given by the computed molecular surface area, sum of charges on the carbon atoms of the chromone moiety, and net atomic charge at 3′ and 4′ position of the phenyl ring. It is gratifying to observe that statistical parameters of both classes of descriptors corroborate nicely in terms of possible mode of binding to the enzymatic active site. Also, comparison of the counterpropagation ANN with more frequently used classical approaches proved that the former is a valuable tool in search for quantitative relationships in biology-related complex systems.

ACKNOWLEDGMENT

Our thanks are due to Professors A. Krbavcic (Pharmaceutical Faculty, Ljubljana) and J. Zupan (National Institute of Chemistry Ljubljana) for their keen interest in the present work and valuable discussions.

REFERENCES AND NOTES

(1) Ullrich, A.; Schlessinger, J. Signal Transduction by Receptors with Tyrosine Kinase Activity. *Cell* **1990**, *61*, 203−212.
(2) Bishop, M. J. The Molecular Genetics of Cancer. *Science* **1987**, *235*, 305−311.
(3) Levitzki, A.; Gazit, A. Tyrosine Kinase Inhibition: An Approach to Drug Development. *Science* **1995**, *267*, 1782−1788.
(4) Burke, R. T. Protein-Tyrosine kinase inhibitors. *Drugs of the Future* **1992**, *17*, 119−131.
(5) Chang, C.; Geahlen, L. R. Protein-tyrosine kinase inhibition:mechanism-based discovery of antitumor agents. *J. Nat. Products* **1992**, *55*(11), 1529−1560.
(6) Groundwater, P. W.; Solomons, H. R. K.; Drewe, A. J.; Munawar, A. M. Protein Tyrosine Kinase Inhibitors. *Prog. Med. Chem.* **1996**, *33*, 233−329.
(7) Pathak, D., Pathak, K.; Singla, A. K. Flavonoids as medicinal agents - Recent advances. *Fitoterapia* **1991**, *62*, 371−389.
(8) Cushman, M.; Nagarathnam, D.; Geahlen, L. R. Synthesis and Evaluation of Hydroxylated Flavones and Related Compounds as Potential Inhibitors of the Protein-Tyrosine Kinase p56. *J. Nat. Products* **1991**, *54*, 1345−1352.
(9) Cushman, M.; Nagarathnam, D.; Burg, L. D.; Geahlen, L. R. Synthesis and Protein-Tyrosine Kinase Inhibitory Activities of Flavonoid Analogues. *J. Med. Chem.* **1991**, *34*, 798−806.
(10) Cushman, M.; Zhu, H.; Geahlen, L. R.; Kraker, J. A. Synthesis and Biochemical Evaluation of a Series of Aminoflavones as Potential Inhibitors of Protein-Tyrosine Kinases p56, EGFr, p60. *J. Med. Chem.* **1994**, *37*, 3353−3362.
(11) Hagiwara, M.; Inoue, S.; Tanaka, T.; Nunoki, K.; Ito, M.; Hidaka, H. Differential Effects of Flavonoids as Inhibitors of Tyrosine Protein Kinases and Serine/Threonine Protein Kinases. *Biochem. Pharmacol.* **1988**, *37*, 2987−2992.
(12) Nikolovska-Coleska, Z., Suturkova, Lj., Dorevski, K., Krbavcic, A., and Solmajer, T. QSAR of Flavonoid Inhibitors of $p56^{lck}$. Protein Tyrosine Kinase: A Quantum Chemical/Classical Approach. submitted.
(13) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Verlag: Weinheim, FRG, 1993.
(14) Rumelhart, E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Back-propagation. In *Distributed Parallel Processing*: *Explorations in the Microstructures of Cognition*; Rumelhart, D. E., MacClelland, J. L., Eds.; MIT Press: Cambridge, MA, 1986; Vol 1, pp 318−362.
(15) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure-Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(16) Andrea, T. A.; Kalayeh, H. Application of Neural Networks in Quantitative Structure-Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(17) Kyngas, J.; Valjakka, J. Evolutionary Neural Networks in Quantitative Structure-Activity Relationship of Dihydrofolate Reductase Inhibitors. *Quant. Struct.-Act. Relat.* **1996**, *15*, 296−301.
(18) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer Verlag: Berlin, FRG, 1989.
(19) Hecht-Nielsen, R. Application of Counter-propagation Networks, *Neural Networks* **1988**, *1*, 131−140.
(20) Novič, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counter-Propagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.
(21) Bienfait, B. Application of High-Resolution Self-Organizing Maps to Retrosynthetic and QSAR Analysis. *J. Chem. Inf. comput. Sci.* **1994**, *34*(4), 890−898.
(22) Peterson, K. L. Quantitative Structure-Activity Relationship in Carboquinones and Benzodiazepines. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896−904.
(23) Hansch, L.; Leo, A. *Exploring QSAR Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.
(24) Spartan User's Guide, Version 4.0; Wavefunction, Inc.: Irvine, CA, 1995.
(25) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
(26) Solmajer, T.; Lukovits, I.; Hadzi, D. *J. Med. Chem.* **1982**, *25*, 1413−1417.
(27) Solmajer, T.; Lukovits, I.; Hadzi, D. *Quant. Struct.-Act. Relat.* **1984**, *3*, 51−59.
(28) Efron, B.; Tibshirani, R. J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.

CI970222P