Modeling of Cyclin-Dependent Kinase Inhibition by 1*H*-Pyrazolo[3,4-*d*]Pyrimidine Derivatives Using Artificial Neural Network Ensembles

Michael Fernández,[†] Alain Tundidor-Camba,[‡] and Julio Caballero*,[†]

Molecular Modeling Group, Center for Biotechnological Studies, University of Matanzas, Matanzas, Cuba, and Scientific Prospection Group, National Centre for Scientific Researches (CNIC), P.O. Box 6880, Havana, Cuba

Received June 25, 2005

Artificial neural network ensembles were used for modeling the cyclin-dependent kinase inhibition of 1*H*-pyrazolo[3,4-*d*]pyrimidine derivatives. The structural characteristics of these inhibitors were encoded in relevant 3D-spatial descriptors extracted by genetic algorithm feature selection. Bayesian-regularized multilayer neural networks, trained by the back-propagation algorithm, were developed using these variables as inputs. The predictive power of the model was tested by leave-one-out cross validation. In addition, for a more rigorous measure of the predictive capacity, multiple validation sets were randomly generated as members of neural network ensembles, which makes doing averaged predictions feasible. In this way, the predictive power was analyzed accounting for the averaged test set *R* values and test set mean-square errors. Otherwise, Kohonen self-organizing maps were used as an additional tool for the same modeling. The location of the inhibitors in a map facilitates the analysis of the connection between compounds and serves as a useful tool for qualitative predictions.

1. INTRODUCTION

The development of cancer therapeutics in recent years has taken on new dimensions since modern biological techniques open the way to understanding key cellular processes at the individual protein level.[1] Cancer is not a single disease but a broad group characterized by uncontrolled proliferative growth and the spread of aberrant cells from their site of origin. At the simplest level, cancer cells may be regarded as having lost touch with their environment so that they are no longer responsive to the controlling signals and interactions which occur continuously in normal, healthy tissues.

Most anticancer agents which have been approved for clinical use are molecules that damage deoxyribonucleic acid (DNA), block DNA synthesis by inhibiting the nucleic acid precursors' biosynthesis, alter tubulin polymerization−depolymerization, or disrupt the hormonal stimulation of cell growth.[2] However, there has been a recent shift of emphasis toward novel mechanistic targets that have directly arisen from the in-depth study of the underlying genetic changes related to the cancerous state. Thus, for instance, the identification of oncogenes such as ras protein, which are activated in tumors, has suggested many pathways and specific molecular entities as rational targets for anticancer drug discovery.[1]

Controlling the cell cycle by inhibition of the proteins that regulate its progression is an attractive strategy for addressing cancer and other diseases associated with abnormal cellular proliferation. Cyclin-dependent kinases (CDKs) constitute a class of serine−threonine (S−T) protein kinases that, in association with specific regulatory subunits (cyclins), play an important role in regulation of the cell cycle.[3] Abnormal CDK control of the cell cycle has been strongly linked to the molecular pathology of cancer; for this reason, CDKs have, thus, become attractive therapeutic targets for cancer therapy.[4] One of the prime targets is CDK4, which acts at the G1/S interface when it is associated to cyclin D1. Passage through an initial G1 restriction point occurs upon release by the retinoblastoma protein (pRb) of the transcription factor E2F.[5] This is triggered by phosphorylation of several S/T residues on pRb, the primary substrate of CDK4/cyclin D. Therefore, inhibition of CDK4 should block entry into the cell cycle; as a result, the fate of cells that do not progress through the G1 or subsequent checkpoints is to undergo apoptosis.

Actually, several CDK inhibitors (CKIs) have entered clinical evaluation for the treatment of cancer. These include flavopiridol,[6] 7-hydroxystaurosporine (UCN-01), roscovitine (CYC202),[7] and the aminothiazole compound (BMS-387032).[8] The identified CKIs display antimitotic and apoptosis-inducing properties and are being evaluated as potential antitumor agents.[9] Furthermore, when used in combination with cytotoxic drugs, they can re-establish an otherwise deficient cell cycle checkpoint, which may enhance the activity of conventional treatments. Nowadays, the synthesis of novel highly selective CKIs as candidates for CDK-target therapy in cancer treatment is in high demand.[10,11]

A broad range of information is available about CDK inhibitors. Until now, however, very few studies on the relationship between the chemical structures and the biological functions of this kind of compound have been reported. As far as we know, the work of Pies et al. is the sole report on a 2D-QSAR analysis of CDK inhibitors (QSAR = quantitative structure−activity relationship).[12] This group

* Corresponding author tel.: (53) (45) 26 1251; fax: (53) (45) 25 3101; e-mail: jmcr77@yahoo.com.
[†] University of Matanzas.
[‡] National Centre for Scientific Researches.

10.1021/ci050263i CCC: $30.25
Published on Web 10/15/2005

employed multiple linear regression analysis for establishing a QSAR model for the CDK1-inhibitory activity of a series of 22 9-substituted paullones.

In this paper, a detailed correlation study was accomplished by nonlinear 2D-QSAR analysis using artificial neural networks (ANNs). ANNs arose from attempts to model the functioning of the human brain.[13] In chemistry and related fields of research like biochemistry, chemical engineering, and pharmacy, interest in ANN computing has grown rapidly. In the past decade, ANNs have encountered successful applications in QSAR studies. ANNs usually overcome methods limited to linear regression models such as multiple linear analysis or partial least square.[14,15] Contrary to these methods, ANNs can be used to model complex nonlinear relationships. Since biological phenomena are complex by nature, this ability has promoted the employment of ANNs in drug design studies.

In this work, the main objective is to model the biological activity of cyclin-dependent kinase inhibitors recently reported by Markwalder et al.[16] The data set consists of 98 1*H*-pyrazolo[3,4-*d*]pyrimidine derivatives with the biological activity reported as IC_{50} values. The characteristics of the inhibitors were represented by relevant 3D descriptors extracted by genetic algorithm (GA) feature selection. Two different ANN models were used in this study: Bayesian-regularized artificial neural networks (BRANNs) trained by a back-propagation algorithm and self-organizing maps (SOMs) trained by competitive learning. This combination exploits the specific characteristics of these two different methods in analyzing and modeling data on biological activities. Both methods can be analyzed together; BRANNs were applied in order to predict IC_{50} values, while SOMs allot similar locations to similar compounds in a feature map.

2. METHODS AND EXPERIMENTAL PROCEDURE

2.1. Data Sets and Feature Selection. CDK4/cyclin D1 inhibitory activities [$\log(10^6/IC_{50})$] of 98 1*H*-pyrazolo[3,4-*d*]pyrimidine derivatives were taken from the literature.[16] The chemical structures and experimental activities are shown in Table 1. IC_{50} values represent the compound concentration that inhibits 50% of the kinase activity. Prior to molecular descriptor calculations, 3D structures of the studied compounds were geometrically optimized using the semiempirical quantum-chemical method PM3[17] implemented in the MOPAC 6.0[18] computer software.

The 3D descriptors from the Dragon software[19] were calculated for each compound: aromaticity indices,[20,21] Randic molecular profiles,[22] geometrical descriptors,[23] RDF descriptors,[24] 3D-MoRSE (molecule representation of structures based on electron diffraction) descriptors,[25] WHIM (weighted holistic invariant molecular) descriptors,[26] and GETAWAY descriptors.[27] In all, 721 descriptors were calculated. Descriptors that stayed constant or almost constant were eliminated, and pairs of variables with a correlation coefficient greater than 0.95 were classified as intercorrelated, and only one of these was included in the model. Finally, 276 descriptors were obtained.

Since many molecular descriptors were available for QSAR analysis and only a reduced subset of them is statistically significant in terms of correlation with biological activities, deriving an optimal QSAR model through variable selection needs to be addressed. Six variables are adequate to describe the dataset of 98 compounds (about 15 molecules per one fitted parameter in the model). As a consequence, a nonlinear GA search was carried out for extracting the six most relevant variables.

GAs are governed by biological evolution rules.[28] They are stochastic optimization methods that have been inspired by evolutionary principles. The distinctive aspect of a GA is that it investigates many possible solutions simultaneously, each of which explores different regions in parameter space.[29] The first step is to create a population of *N* individuals. Each individual encodes the same number of randomly chosen descriptors. The fitness of each individual in this generation is determined. In the second step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents on the basis of their scaled fitness scores. The new offspring contain characteristics from two or one of its parents.

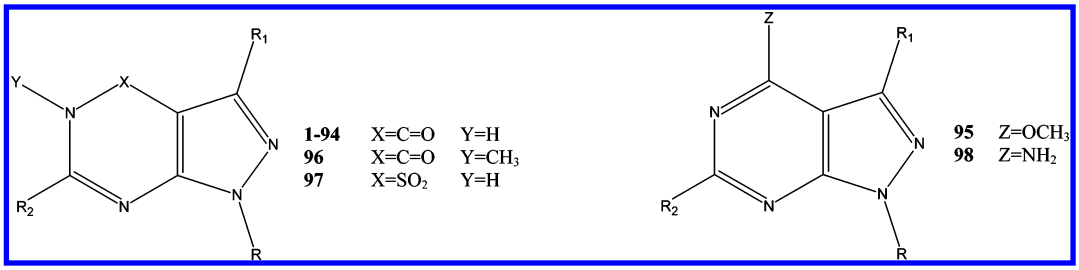
The GA implemented in this paper is a version of the So and Karplus report[30] and was programmed within the MATLAB environment using the genetic algorithm and neural networks toolboxes.[31]

In our approach, the predictors are BRANNs with a 6−2−1 architecture, and the mean-square error (MSE) of data fitting was tried as the individual fitness function. An individual in the population is represented by a string of integers which reflects the numbering of the columns in the data matrix. In the So and Karplus report,[30] the fitness of the individual was determined by a variety of fitness functions which are proportional to the residual error of the training set, the test set, or even the cross-validation set from the neural network simulations. In our approach, we tried the MSE of data fitting for BRANN models, as the case may be, as the individual fitness function. The first step is to create a gene pool (population) of *N* individuals. Each individual encodes the same number of descriptors; the descriptors are randomly chosen from a common data matrix, and in a way such that (1) no two individuals can have exactly the same set of descriptors and (2) all descriptors in a given individual must be different. The fitness of each individual in this generation is determined by the MSE of the model and scaled using a scaling function. A top scaling fitness function scaled a top fraction of the individuals in a population equally; these individuals have the same probability to be reproduced, while the rest are assigned the value 0.

In the next step, a fraction of children of the next generation is produced by crossover (crossover children) and the rest by mutation (mutation children) from the parents. Sexual and asexual reproductions take place so that the new offspring contain characteristics from two or one of its parents. In a sexual reproduction, two individuals are selected probabilistically on the basis of their scaled fitness scores and serve as parents. Next, in a crossover, each parent contributes a random selection of half of its descriptor set and a child is constructed by combining these two halves of "genetic code". Finally, the rest of the individuals in the new generation are obtained by asexual reproduction when parents selected randomly are subjected to a random mutation in one of their genes; that is, one descriptor is replaced by another.

Similarly to So and Karplus,[30] we also included elitism, which protects the fittest individual in any given generation

Table 1. Experimental and Predicted CDK4/Cyclin D1 Inhibitory Activities of 1*H*-Pyrazolo[3,4-*d*]pyrimidine Derivatives



compound	*R*	R_1	R_2	exptl $\log(10^6/IC_{50})$	training predictions	external predictions
1	2,6-dichloro-4-(CF_3)phenyl	CH_3S-	CH_3	5.10	4.77	4.35
2	benzyl	CH_3S-	CH_3	3.82	4.05	4.33
3	*n*-butyl	CH_3S-	CH_3	3.70	3.57	3.69
4	*tert*-butyl	CH_3S-	CH_3	4.00	3.92	4.00
5	2-hydroxyethyl	CH_3S-	CH_3	3.68	3.79	3.79
6	phenyl	CH_3S-	CH_3	3.74	3.60	3.53
7	2-pyridyl	CH_3S-	CH_3	3.74	3.90	3.94
8	3-chlorophenyl	CH_3S-	CH_3	3.80	3.95	3.90
9	4-chlorophenyl	CH_3S-	CH_3	3.80	3.69	3.63
10	4-isopropylphenyl	CH_3S-	CH_3	3.80	3.87	3.99
11	2-methoxyphenyl	CH_3S-	CH_3	4.08	4.06	4.25
12	2-chlorophenyl	CH_3S-	CH_3	4.72	4.20	4.19
13	4-chloro-2-methylphenyl	CH_3S-	CH_3	4.13	4.48	4.50
14	2,3-dichlorophenyl	CH_3S-	CH_3	4.30	4.37	4.50
15	2,4-dimethylphenyl	CH_3S-	CH_3	4.32	4.27	4.20
16	2,4-dichlorophenyl	CH_3S-	CH_3	4.43	4.27	4.19
17	2,5-dimethylphenyl	CH_3S-	CH_3	4.60	4.45	4.23
18	2,5-dichlorophenyl	CH_3S-	CH_3	4.64	4.72	4.69
19	2-chloro-5-(CF_3)phenyl	CH_3S-	CH_3	4.82	4.80	4.39
20	2-chloro-6-fluorophenyl	CH_3S-	CH_3	5.17	4.95	4.77
21	2,6-dichlorophenyl	CH_3S-	CH_3	5.28	4.86	4.84
22	2,4-dichloro-6-(CF_3)phenyl	CH_3S-	CH_3	4.54	4.87	4.64
23	2,4-dichloro-6-methylphenyl	CH_3S-	CH_3	4.96	4.50	4.35
24	2,4,6-trichlorophenyl	CH_3S-	CH_3	5.68	4.74	4.61
25	2,4,6-trimethylphenyl	CH_3S-	CH_3	4.43	4.37	4.41
26	2,4,6-trichlorophenyl	CN	CH_3	4.55	4.28	4.27
27	2,4,6-trichlorophenyl	CF_3	CH_3	5.51	4.98	4.84
28	2,4,6-trichlorophenyl	phenyl	CH_3	3.92	4.11	4.47
29	2,4,6-trichlorophenyl	benzyl	CH_3	4.32	4.50	4.59
30	2,4,6-trichlorophenyl	*n*-butyl	CH_3	5.19	5.52	5.71
31	2,4,6-trichlorophenyl	*n*-propyl	CH_3	5.24	5.61	5.62
32	2,4,6-trichlorophenyl	C_2H_5	CH_3	5.96	5.54	5.39
33	2,4,6-trichlorophenyl	CH_3	CH_3	5.46	5.16	5.09
34	2,4,6-trichlorophenyl	H	CH_3	4.92	5.11	5.06
35	2,4,6-trichlorophenyl	$HOCH_2$-	CH_3	5.15	4.68	4.66
36	2,4,6-trichlorophenyl	$HOCH_2CH_2$-	CH_3	3.85	4.33	4.59
37	2,4,6-trichlorophenyl	CH_3S-	H	4.02	4.81	4.94
38	2,4,6-trichlorophenyl	CH_3S-	isopropyl	5.00	4.73	4.82
39	2,4,6-trichlorophenyl	CH_3S-	*n*-propyl	5.04	4.82	4.91
40	2,4,6-trichlorophenyl	CH_3S-	cyclopropyl	5.37	4.82	4.73
41	2,4,6-trichlorophenyl	CH_3S-	BocN(Me)CH_2-	4.30	4.43	4.26
42	2,4,6-trichlorophenyl	CH_3S-	BocN(Me)CH_2CH_2-	4.32	4.38	4.25
43	2,4,6-trichlorophenyl	CH_3S-	$CH_3CH(OH)$-	4.57	4.85	5.05
44	2,4,6-trichlorophenyl	CH_3S-	CH_3SCH_2-	4.82	4.61	4.77
45	2,4,6-trichlorophenyl	CH_3S-	MeO_2C-	3.62	3.79	5.13
46	2,4,6-trichlorophenyl	CH_3S-	$HO(CH_2)_3$-	4.89	4.83	4.67
47	2,4,6-trichlorophenyl	CH_3S-	$HO(CH_2)_5$-	5.00	4.88	5.31
48	2,4,6-trichlorophenyl	CH_3S-	CH_2F	4.36	4.72	4.84
49	2,4,6-trichlorophenyl	CH_3S-	CF_3	3.94	4.29	4.49
50	2,4,6-trichlorophenyl	CH_3S-	2-furyl	4.44	4.84	4.79
51	2,4,6-trichlorophenyl	CH_3S-	CH_2−2-thienyl	4.70	4.84	4.99
52	2,4,6-trichlorophenyl	CH_3S-	benzyl	4.96	4.77	4.83
53	2,4,6-trichlorophenyl	C_2H_5	isobutyl	5.20	5.24	5.27
54	2,4,6-trichlorophenyl	C_2H_5	C_2H_5	6.03	5.98	5.73
55	2,4,6-trichlorophenyl	C_2H_5	$CH_3SCH_2CH_2$-	5.19	4.99	5.09
56	2,4,6-trichlorophenyl	C_2H_5	$HO(CH_2)_4$-	5.62	5.85	5.58
57	2,4,6-trichlorophenyl	C_2H_5	3,4-dimethoxyphenyl	4.68	4.88	5.14
58	2,4,6-trichlorophenyl	C_2H_5	CH_2CH_2-phenyl	5.15	5.25	5.14
59	2,4,6-trichlorophenyl	C_2H_5	CH_2CH_2-imidazol-4-yl	5.44	5.73	5.89
60	2,4,6-trichlorophenyl	C_2H_5	benzyl	5.52	5.68	5.73
61	2,4,6-trichlorophenyl	C_2H_5	2-methoxy benzyl	4.66	4.80	4.91

Table 1 (Continued)

compound	*R*	R_1	R_2	exptl $\log(10^6/IC_{50})$	training predictions	external predictions
62	2,4,6-trichlorophenyl	C_2H_5	2-(hydroxymethyl) benzyl	5.54	5.74	5.34
63	2,4,6-trichlorophenyl	C_2H_5	2-pyridinylmethyl	5.01	4.70	4.81
64	2,4,6-trichlorophenyl	C_2H_5	3-(amino-2-methyl) benzyl	5.66	5.87	6.11
65	2,4,6-trichlorophenyl	C_2H_5	3-methyl benzyl	5.66	5.63	5.52
66	2,4,6-trichlorophenyl	C_2H_5	3-(ethoxycarbonylmethyl) benzyl	5.72	5.92	6.69
67	2,4,6-trichlorophenyl	C_2H_5	3-methoxy benzyl	5.30	5.65	5.71
68	2,4,6-trichlorophenyl	C_2H_5	3-amino benzyl	5.54	5.69	5.95
69	2,4,6-trichlorophenyl	C_2H_5	3-pyridinylmethyl	5.89	5.70	5.58
70	2,4,6-trichlorophenyl	C_2H_5	4-pyridinylmethyl	5.52	5.71	5.66
71	2,4,6-trichlorophenyl	C_2H_5	4-amino benzyl	5.57	5.68	5.81
72	2,4,6-trichlorophenyl	C_2H_5	4-methoxy benzyl	5.66	6.30	6.36
73	2,4,6-trichlorophenyl	C_2H_5	4-hydroxy benzyl	5.70	5.70	5.92
74	2,4,6-trichlorophenyl	C_2H_5	4-bromo benzyl	5.92	5.78	5.71
75	2,4,6-trichlorophenyl	C_2H_5	4-dimethylamino benzyl	5.68	5.46	5.27
76	2,4,6-trichlorophenyl	C_2H_5	4-methoxy-3-methyl benzyl	5.68	5.55	5.43
77	2,4,6-trichlorophenyl	C_2H_5	3-methoxy-4-methyl benzyl	5.68	5.86	5.90
78	2,4,6-trichlorophenyl	C_2H_5	4-hydroxy-3-methoxy benzyl	5.18	5.16	5.11
79	2,4,6-trichlorophenyl	C_2H_5	3,4-dimethoxy benzyl	5.35	5.38	5.88
80	2,4,6-trichlorophenyl	C_2H_5	3,5-dihydroxy benzyl	7.23	7.11	7.04
81	2,4,6-trichlorophenyl	C_2H_5	3-hydroxy-4-methoxy benzyl	7.62	7.28	6.76
82	2,4,6-trichlorophenyl	CH_3S-	$MeNHCH_2$·HCl	4.24	4.78	5.05
83	2,4,6-trichlorophenyl	CH_3S-	$MeNHCH_2CH_2$·HCl	4.55	4.39	4.22
84	2,4,6-trichlorophenyl	C_2H_5	vinyl	5.54	5.25	5.12
85	2,4,6-trichlorophenyl	C_2H_5	2-hydroxy benzyl	4.96	5.15	5.40
86	2,4,6-trichlorophenyl	C_2H_5	3-hydroxy benzyl	7.17	6.43	6.14
87	2,4,6-trichlorophenyl	C_2H_5	4-hydroxy-3-methyl benzyl	5.66	5.95	6.11
88	2,4,6-trichlorophenyl	C_2H_5	3-hydroxy-4-methyl benzyl	7.06	6.24	6.13
89	2,4,6-trichlorophenyl	C_2H_5	3,4-dihydroxy benzyl	7.36	6.83	6.42
90	2,4,6-trichlorophenyl	C_2H_5	3-methylsulfonamido benzyl	6.80	6.86	6.67
91	2,4,6-trichlorophenyl	C_2H_5	4-(methylcarbonyl) benzyl	5.89	6.00	5.82
92	2,4,6-trichlorophenyl	C_2H_5	4-($(CH_3)_3COCONH-CH_2CO$) benzyl	5.77	5.79	5.98
93	2,4,6-trichlorophenyl	C_2H_5	4-(H_2NCH_3CO·HCl) benzyl	6.34	6.19	5.87
94	2,4,6-trichlorophenyl	C_2H_5	4-($(CH_3)_2NCH_2CO$·HCl) benzyl	6.23	6.10	6.04
95	2,4,6-trichlorophenyl	C_2H_5	CH_3	4.57	4.53	4.55
96	2,4,6-trichlorophenyl	C_2H_5	CH_3	4.57	4.84	4.88
97	2,4,6-trichlorophenyl	C_2H_5	CH_3	4.29	5.13	5.66
98	2,4,6-trichlorophenyl	C_2H_5	CH_3	3.85	4.21	4.46

from crossover or mutation during reproduction. The genetic content of this individual simply moves intact on to the next generation. This selection, crossover, and mutation process is repeated until all of the *N* parents in the population are replaced by their children. The fitness score of each member of this new generation is again evaluated, and the reproductive cycle is continued until 90% of the generations showed the same target fitness score.[32] The best models according to *R* value ($R > 0.8$) were selected, and they were tested in cross-validation experiments for avoiding chance correlations.

2.2. Bayesian-Regularized Artificial Neural Networks. ANNs are computer-based models in which a number of processing elements, also called neurons, units, or nodes, are interconnected by links in a netlike structure forming "layers".[13,33] Every connection between two neurons is associated with a weight, a positive or negative real number that multiplies the signal from the preceding neuron. Neurons are commonly distributed among the input, hidden, and output layers. Neurons in the input layer receive their values from independent variables; in turn, the hidden neurons collect values from precedent neurons, giving a result that is passed to a successor. Finally, neurons in the output layer take values from other units and correspond to different dependent variables.

Commonly, ANNs are adjusted, or trained, so that a particular input leads to a specific target output. According to this, the output *j* is obtained from the input *j*, by application of eq 1:

$$\mathrm{out}_j = f(\mathrm{inp}_j) \tag{1}$$

where the function *f* is called a transfer function. When the ANN is training, the weights are updated in order to minimize network error. In contrast to common statistical methods, ANNs are not restricted to linear correlations or linear subspaces.[13] The employed transfer function, commonly, a hyperbolic tangent function, allows the establishment of nonlinear relations. Thus, ANNs can take into account nonlinear structures and structures of arbitrarily shaped clusters or curved manifolds.

While more connections take effect, the ANN better adjusts the input−output relation. However, when parameters increase, the network loses its ability to generalize. The error on the training set is driven to a very small value, but when new data is presented to the network, the error is large. In this process, the predictor has memorized the training examples, but it has not learned to generalize to new situations; it means network overfits the data.

Typically, training aims to reduce the sum of squared errors:

$$F = \mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2 \tag{2}$$

In this equation, F is the network performance function, MSE is the mean of the sum of squares of the network errors, N is the number of compounds, y_i is the predicted biological activity of compound i, and t_i is the experimental biological activity of compound i.

MacKay's Bayesian-regularized ANNs (BRANNs) have been designed to resist overfitting.[34] To accomplish this purpose, BRANNs include an error term that regularizes the weights by penalizing overly large magnitudes.

Assuming a set of pairs $D = \{x_i, t_i\}$, where $i = 1-N$ is a label running over the pairs, the data set can be modeled as deviating from this mapping under some additive noise process (υ_i):

$$t_i = y_i + \upsilon_i \tag{3}$$

If υ is modeled as zero-mean Gaussian noise with standard deviation σ_υ, then the probability of the data, given the parameters w, is

$$P(D|w,\beta,M) = \frac{1}{Z_D(\beta)} \exp(-\beta \times \mathrm{MSE}) \tag{4}$$

where M is the particular neural network model used, $\beta = 1/\sigma_\upsilon^2$, and the normalization constant is given by $Z_D(\beta) = (\pi/\beta)^{N/2}$. $P(D|w,\beta,M)$ is called the likelihood. The maximum likelihood parameters w_{ML} (the w that minimizes MSE) depend sensitively on the details of the noise in the data.

For completing the interpolation model, a prior probability distribution must be defined which embodies our prior knowledge on the sort of mappings that are "reasonable".[35] Typically, this is quite a broad distribution, reflecting the fact that we only have a vague belief in a range of possible parameter values. Once we have observed the data, Bayes' theorem can be used to update our beliefs, and we obtain the posterior probability density. As a result, the posterior distribution is concentrated on a smaller range of values than the prior distribution. Since a neural network with large weights will usually give rise to a mapping with a large curvature, we favor small values for the network weights. At this point, a prior value is defined that expresses the sort of smoothness that the interpolant is expected to have. The model has the form

$$P(w|\alpha,M) = \frac{1}{Z_W(\alpha)} \exp(-\alpha \times \mathrm{MSW}) \tag{5}$$

where α represents the inverse variance of the distribution and the normalization constant is given by $Z_W(\alpha) = (\pi/\alpha)^{N/2}$. MSW is the mean of the sum of the squares of the network weights and is commonly referred to as a regularizing function.

Considering the first level of inference, if α and β are known, then the posterior probability of the parameters w is where $P(w|D,\alpha,\beta,M)$ is the posterior probability, that is, the

$$P(w|D,\alpha,\beta,M) = \frac{P(D|w,\beta,M) \times P(w|\alpha,M)}{P(D|\alpha,\beta,M)} \tag{6}$$

plausibility of a weight distribution considering the information of the data set in the model used; $P(w|\alpha,M)$ is the prior density, which represents our knowledge of the weights before any data is collected; $P(D|w,\beta,M)$ is the likelihood function, which is the probability of the data occurring, given the weights; and $P(D|\alpha,\beta,M)$ is a normalization factor, which guarantees that the total probability is 1.

Considering that the noise in the training set data is Gaussian and that the prior distribution for the weights is Gaussian, the posterior probability fulfills the relation:

$$P(w|D,\alpha,\beta,M) = \frac{1}{Z_F} \exp(-F) \tag{7}$$

where Z_F depends on objective function parameters. So, under this framework, the minimization of F is identical to finding the (locally) most probable parameters.[34]

In short, Bayesian regularization involves modifying the performance function (F) defined in eq 2, which is possible when improving generalization by adding an additional term.

The relative size of the objective function parameters α

$$F = \beta \times \mathrm{MSE} + \alpha \times \mathrm{MSW} \tag{8}$$

and β dictates the emphasis for getting a smoother network response. MacKay's Bayesian framework automatically adapts the regularization parameters to maximize the evidence of the training data.[34]

Bayesian regularization overcomes the remaining deficiencies of neural networks and produces predictors that are robust and well-matched to the data; in this sense, BRANNs have been successfully applied in QSAR analysis.[36]

Fully connected, three-layer BRANNs with back-propagation training were implemented in the MATLAB environment.[31] In these nets, the transfer functions of input and output layers were linear and the hidden layer had neurons with a hyperbolic tangent transfer function. Inputs and targets took the values from independent variables selected by the GA and $\log(10^6/\mathrm{IC}_{50})$ values, respectively; both were normalized prior to network training. BRANN training was carried out according to the Levenberg−Marquardt optimization.[37] The initial value for μ was 0.005 with decrease and increase factors of 0.1 and 10, respectively. The training was stopped when μ became larger than 10^{10}.

2.3. Artificial Neural Network Ensembles. An artificial neural network ensemble (NNE) is a learning paradigm where many ANNs are jointly used to solve a problem. On the basis of this judgment, a collection of a finite number of neural networks is trained for the same task and the outputs can be combined to form one unified prediction. As a result, the generalization ability of the neural network system can be significantly improved.[38]

An effective NNE should consist of a set of ANNs that not only are highly correct but make their errors on different parts of the input space as well. So, the combination of the output of several classifiers is only useful if they disagree on some inputs. Krogh and Vedelsby[39] later proved that the ensemble error can be divided into a term measuring the average generalization error of each individual network and a term called diversity that measures the disagreement among the networks. Formally, they defined the diversity term d_i of network i on input j to be

$$d_i(j) = [o_i(j) - \bar{o}(j)]^2 \tag{9}$$

where $o_i(j)$ and $\bar{o}(j)$ are the ith classifier and the ensemble predictions, respectively. In other words, it is simply the

variance of the ensemble around the mean. The quadratic errors of network i and those of the ensemble are, respectively,

$$\epsilon_i(j) = [o_i(j) - f(j)]^2 \tag{10}$$

and

$$e(j) = [\bar{o}(j) - f(j)]^2 \tag{11}$$

where $f(j)$ is the target value for input j. If we define $\bar{E}$, E_i, and D_i to be the averages, over the input distribution of $e(j)$, $\epsilon(j)$, and $d(j)$, respectively, then the ensemble's generalization error can be shown to consist of two distinct portions:

$$\bar{E} = E - D \tag{12}$$

where E ($= \sum_i w_i E_i$) is the weighted average of the individual networks' generalization error and D ($= \sum_i w_i D_i$) is the weighted average of the diversity among these networks. What eq 12 shows, then, is that an ideal ensemble consists of highly correct ANNs that disagree as much as possible. In this way, the mean-square error (MSE $= \bar{E}$) of the ensemble estimator is guaranteed to be less than or equal to the averaged mean-square error of the component estimators.

2.4. Model Validation. QSAR has been traditionally perceived as a means of establishing correlations between trends in chemical structure modifications and respective changes of biological activity. Of course, any QSAR modeling should ultimately lead to statistically robust models capable of making accurate and reliable predictions of biological activities of compounds, placing a special emphasis on the statistical significance and predictive ability of these models as their most crucial characteristics.[40] The basis for any QSAR model is that the biological activity of a new or untested chemical can be inferred from the molecular structure or properties of similar compounds whose activities have already been assessed. The statistical fit of a QSAR can be assessed in many easily available statistical terms (e.g., R^2, S, etc.), but this is not strictly related to the ability of a model to make predictions.

The traditional method to obtain a meaningful assessment of statistical fit consists of predicting some removed certain proportion of the data set. The whole data is randomly split into a number of disjointed cross-validation subsets. One of each of these subsets is left out in turn, and the remaining complement of data is used to make a partial model. The samples in the left-out data are then used to perform predictions. At the end of this process, there are predictions for all data in the training set, made up from the predictions originating from the resulting partial models. All partial models are then assessed against the same performance criteria, and decisions are made on the basis of the consistency of the assessment results. The more-used cross-validation method is the leave-one-out (LOO) cross validation, when all cross-validation subsets consist of only one data point each.

The accuracy of cross-validation results is extendedly accepted in the literature considering the Q^2 value. In this sense, a high value of the statistical characteristic ($Q^2 > 0.5$) is considered as proof of the high predictive ability of the model.[41] However, several authors suggest that a high value of Q^2 appears to be a necessary but not sufficient condition for the model to have a high predictive power and consider that the predictive ability of a QSAR model can only be estimated using a sufficiently large collection of compounds that was not used for building the model.[42] In this sense, the data set can be divided into training and validation (or test) partitions. For the given partitioning, a model is constructed only of samples of the training set. At this point, an important step is the generation of theses partitions. Quite a few methods have been used, such as random selection, activity-ranked binning, and sphere exclusion algorithms.[43] Various forms of neural networks have also been employed in the selection of training sets, including Kohonen neural networks.[44,45]

Undoubtedly, external validation is a way to establish the reliability of a QSAR model. However, the majority of studies that are validated by external predictions are based on a single validation set; this may cause the predictors to perform well on a particular external set, but there is no guarantee that the same results may be achieved on another. For example, it can happen that several outliers, by pure coincidence, are out of the test set, in which case, the validation error will be small even though the training error was high.

Previously, the ensemble solution has been proposed for originating multiple validation sets.[46] An ensemble is a collection of predictors that, as a whole, provides a prediction which is a combination of the individual ones. If there is disagreement among those predictors, very reliable models can be obtained, since a further decrease in generalization error can be achieved. Another trait to take into account for the ensemble application is the average error of ensemble members; with this trait, when decreasing the error for each individual, the ensemble gets a smaller generalization error.

In this work, the quality of the fit of BRANN models was measured by its R^2 value. Validations were carried out analyzing Q^2 values from LOO cross-validation experiments. In addition, the predictive power of the models was also measured by an external validation process that consists of predicting the activity of unknown compounds forming the test set. To avoid the influence of casual external sets, neural network ensembles are proposed, building all members by the random partition of the whole data set into a training set (80%) and a test set (20%). As a result, averaging external predictions were obtained. The quality and reliability was settled by examining the correlation coefficient R and the MSE of the test set fitting.

2.5. Self-Organizing Maps. Despite back-propagated neural networks having been extensively preferred for nonlinear QSAR modeling, SOMs have also been reported as useful ANNs, accounting for important merits and widespread applications.[47,48]

SOMs[49] are a class of unsupervised neural networks whose characteristic feature is their ability to map nonlinear relations in multidimensional data sets into easily visualizable two-dimensional grids of neurons. SOMs are also referred to as self-organized topological feature maps since the basic function of a SOM is to display the topology of a data set, that is, the relationships between members of the set. These relationships are gathered in several clusters; each local group has the result that topologically close neurons react similarly when they receive similar input. Essentially, SOMs permit the perception of similarities in objects.

In SOMs, the input units are fully connected to the 2D Kohonen layer. Each neuron within the Kohonen layer has a well-defined topology, which means there is a defined number of neurons in its neighborhood. The SOMs are trained through an unsupervised competitive learning process using a "winner takes all" policy. Under this process, molecule *s*, characterized by *m* descriptors x_{si}, will be projected into neuron c_s, which has weights w_{ji}, which are most similar to the input variables (eq 13).

$$\text{out}_{cs} \leftarrow \min[\sum_{i=1}^{m}(X_{si} - wij)^2] \tag{13}$$

Albeit, all neurons in the active layer obtain the same multidimensional input pattern at the same time, only one is selected to represent this pattern. That neuron is avowed as the winner because it has the smallest Euclidian distance between the presented *m*-dimensional input pattern vector x_s (x_{s1}, x_{s2}, ..., x_{si}, ..., x_{sm}) and the *m*-dimensional weight vector w_i (w_{il}, w_{i2}, ..., w_{ij}, ..., w_{im}) of the *i* neurons.

Learning within a Kohonen layer consists of the adjustment of the weights, w_{ij}, in such a manner that the weights of the winning neuron c_s are shifted closer to the values of the input data. However, not only the weights of the winning neuron are adjusted but also those of the neighboring neurons. Equation 14 gives the correction formula for the weights.

$$\Delta w_{ij} = w_{ij} + f(x_{si} - w_{ij}) \tag{14}$$

The correction factor *f* has the largest value for the weights in the winning neuron c_s and decreases with increasing distance between winning and neighboring neurons. Therefore, when a training case is presented to the network, and the winning neuron found, the winner updates its weights using the current learning rate, while the neighbors scale down their weights in proportion to the distance to the winner.

To settle structural similarities among the 1*H*-pyrazolo-[3,4-*d*]pyrimidine derivatives, a Kohonen SOM was built. The 3D descriptors selected by GA were used for unsupervised training of 12 × 12 neuron maps. SOMs were implemented in the MATLAB environment.[31] Neurons were initially located in a grid topology. The ordering phase was developed in 1000 steps with a 0.9 learning rate until a tuning neighborhood distance (1.0) was achieved. The tuning-phase learning rate was 0.02. Training was performed for a period of 2000 epochs in an unsupervised manner.

3. RESULTS AND DISCUSSION

The six variables selected by GA are depicted in Table 2. The vectorial space containing geometrical descriptor *G*(N···O), four 3D-MoRSE descriptors, and the WHIM descriptor E3s has the relevant 3D structural information that models the CDK inhibition in a better way. It is noteworthy that there is no significant intercorrelation between these descriptors, as it is seen in Table 3.

Geometrical descriptor *G*(N···O) accounts for the geometrical distances between N and O atoms. In our data, this descriptor suggests a positive effect of oxygen-containing substituents in position 6 of the 1*H*-pyrazolo[3,4-*d*]pyrimidine derivatives. *G*(N···O) takes a null value only in the poorly active compound **98**, which does not contain oxygen atoms.

Table 2. Symbols of the 3D Descriptors Selected by Genetic Algorithm and Their Definitions

variable	definition
G(N···O)	sum of geometrical distances between N···O
Mor04u	3D MoRSE signal 04 unweighted
Mor30u	3D MoRSE signal 30 unweighted
Mor18m	3D MoRSE signal 18 weighted by atomic masses
Mor11v	3D MoRSE signal 11 weighted by atomic van der Waals volumes
E3s	third component accessibility directional WHIM index weighted by atomic electrotopological states

Table 3. Correlation Matrix of the Descriptors Selected by Genetic Algorithm

	G(N···O)	Mor04u	Mor30u	Mor18m	Mor11v	E3s
G(N···O)	1.000					
Mor04u	0.202	1.000				
Mor30u	0.140	0.000	1.000			
Mor18m	0.011	0.017	0.105	1.000		
Mor11v	0.079	0.001	0.016	0.098	1.000	
E3s	0.071	0.058	0.141	0.001	0.047	1.000

3D-MoRSE[25] code considers the molecular information derived from an equation used in electron diffraction studies. Electron diffraction does not directly yield atomic coordinates but provides diffraction patterns from which the atomic coordinates are derived by mathematical transformations. 3D-MoRSE code is applied by eq 15:

$$I(s) = \sum_{i=2}^{N}\sum_{j=1}^{i-1} A_i A_j \frac{\sin sr_{ij}}{sr_{ij}} \tag{15}$$

In this equation, A_i and A_j are atomic properties of atoms *i* and *j*, r_{ij} represents the interatomic distances, and *s* measures the scattering angle. The value of *s* (0, ..., 31.0 Å^{-1}) is considered only at discrete positions within a certain range. Values of *I*(*s*) are defined at 32 evenly distributed values of *s* in the range of 0−31.0 Å^{-1}. These 32 values constitute the 3D-MoRSE code of the three-dimensional structure of a molecule. Different atomic properties A_i were used, such as atomic mass, atomic van der Waals volumes, residual atomic Sanderson electronegativities, and atomic polarizabilities. The possibility for choosing an appropriate atomic property gives great flexibility to the 3D-MoRSE code for adapting it to the problem under investigation. In this work, 3D-MoRSE-selected descriptors were the unweighted Mor04u and Mor30u, weighted by atomic mass Mor18m and weighted by van der Waals volumes Mor11v. This code represents a restricted 3D space which captures relevant molecular information, regarding molecular size and shape, which is related to the modeled CDK inhibitory activity.

The WHIM indices are invariant to rototraslation descriptors obtained for each molecular geometry.[26] They were calculated by transforming Cartesian coordinates weighted by atomic properties and centering the coordinates to get invariance to translation. Then, a principal component analysis (PCA) leads to three principal component axes, and new coordinates are achieved by projecting the old ones onto the PCA axes, obtaining three score column vectors t_1, t_2, and t_3. Four kinds of descriptors were calculated from the

Table 4. Statistics of the BRANN Models for Cyclin-Dependent Kinase Inhibition of 1*H*-Pyrazolo[3,4-*d*]pyrimidine Derivatives[a]

number of hidden nodes	R^2 [b]	Q^2 [c]	S_{CV} [c]
2	0.796	0.636	0.462
3	0.841	0.595	0.481
4	**0.873**	**0.647**	**0.472**
5	0.897	0.637	0.478
6	0.899	0.604	0.519

[a] Optimum predictor appears in bold. [b] Predicted vs experimental values. [c] LOO cross-validation experiments.

first to fourth order of t_m scores, related to molecular size, shape, symmetry, and atom distribution. The WHIM descriptor E3s is constituted by the kurtosis, calculated from the fourth order moments of the t_m scores weighted by the electrotopological states. It is related to the atom distribution along the third axis for the electrotopological states-weighted scheme. According to this, E3s represents a combination of both electronic and topological characteristics in a mathematically defined zone.

Despite that interpreting QSAR models is always a difficult task, we can conclude that the nonlinear structural information here obtained showed that an adequate distribution of atomic masses, van der Waals atomic volumes, and electrotopological states has a great influence on the CDK4/cyclin D1 inhibitory activities of the studied compounds. It suggests that molecular size, shape, and electronic distribution play an important role in the modeled activity. These facts agree well with reports where the capacity of the active molecules to penetrate the ATP binding site in the monomeric CDK and the facilities to form hydrogen bonds are considered key factors for displaying adequate inhibitory activity.

For an accurate application of BRANN methodology, the network hidden layer's architecture was varied from two to six neurons. The results are depicted in Table 4. The R^2 values were slightly increased with the augmentation of the number of neurons in the hidden layer until a maximum value of about 0.948 was reached. On the other hand, when LOO cross validation was used, Q^2 reached the maximum value of 0.648 when the 6−4−1 architecture was employed. As can be seen, the correlation and predictive power behaviors were satisfactory so that the ability of the GA-selected descriptors to act as relevant information for the modeled activity is also confirmed in this case. A plot of predicted versus experimental $\log(10^6/IC_{50})$ values for the best BRANN model is shown in Figure 1.

A more reliable method for evaluating the predictive power of QSAR models involves predicting activities of unknown compounds. The development of this process requires the selection of a test set. The main characteristic of the test set is that it must contain elements ranging from less- to more-active. The major hitch of this method is that several selected test sets lead to extremely varied results; therefore, a false idea of the predictive power can be achieved if an "overmatched" test set is selected. In this work, we eradicated this trouble by averaging test set predictions using neural network ensembles.

Changing the elements that constitute test and training sets is a way to introduce diversity to the ensemble.[46] In this work, members of NNEs were randomly generated by dividing the whole data set into 78 inhibitors for the training sets and 20

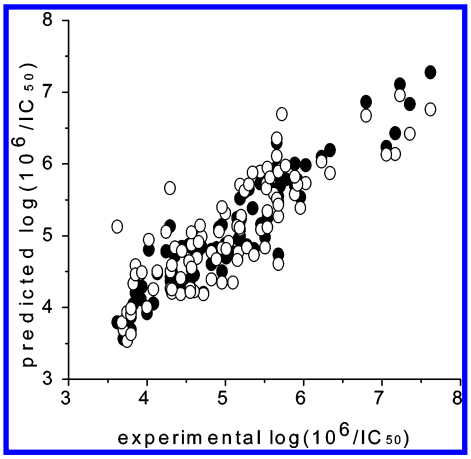


Figure 1. Plot of predicted versus experimental $\log(10^6/IC_{50})$ values for cyclin-dependent kinase inhibition by 1*H*-pyrazolo[3,4-*d*]-pyrimidine derivatives using BRANN models. (●) training predictions; (○) external predictions employing 20-member NNEs.

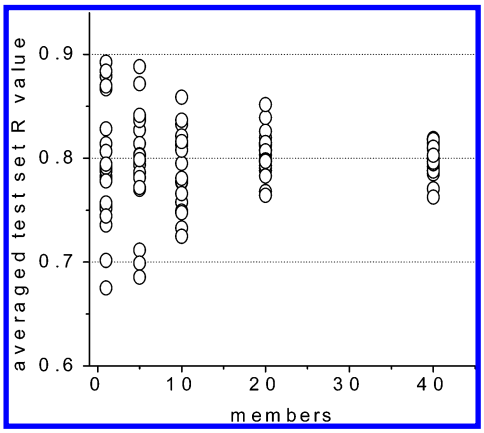


Figure 2. Multiple correlation coefficients (R) of 20 replies of neural network ensembles using 1, 5, 10, 20, and 40 members for test sets.

inhibitors for the test sets, keeping the previous BRANN architecture. Averaged multiple correlation coefficients (averaged R) of the test sets for 20 instances of NNEs with 1, 5, 10, 20, and 40 predictors were examined (Figure 2). All averaged R values stem from adding up 1, 5, 10, 20, or 40 external set R values containing 20 inhibitors each.

NNEs containing one predictor are cases when single training and test sets were selected without integrating them to any ensemble. It is obvious that random partitioning is highly unsatisfactory. Diverse partitions generate a broad scope of external sets, and some are far better predicted than other ones. Therefore, even when it is broadly employed, to assess the predictive power by means of a single external data set, random selection yields a rather fortuitous result. It is noteworthy that more reliable information can be acquired when the number of members in the ensemble is increased. Ensembles containing 40 members are similarly predictive according to averaged test set R values around 0.800. The accumulation of members leads to an averaged model that weights the contribution of each predictor; in this form, deceptive conclusions are suppressed.

Correlation coefficients are good indicators but are not conclusive. More detailed conclusions can be obtained by analyzing the test set MSE values using the Krogh and Vedelsby analysis.[39] For this purpose, the averaged mean

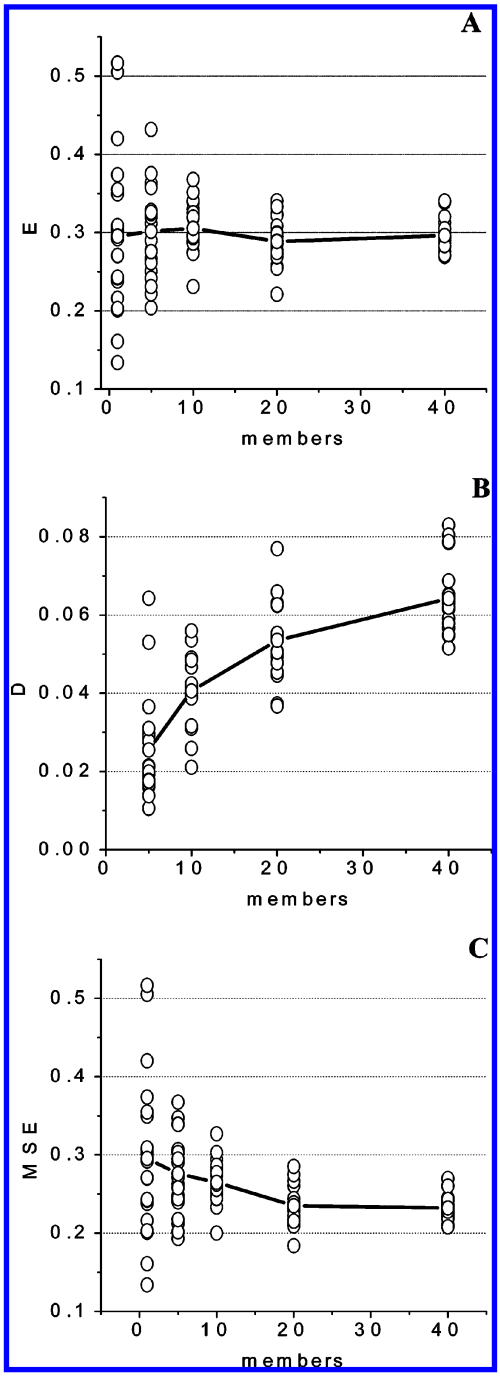


Figure 3. Krogh and Vedelsby analysis for external sets of 20 replies of neural network ensembles using 1, 5, 10, 20, and 40 members. (a) Averaged mean-square errors (*E*). (b) Diversity terms (*D*). (c) Mean-square errors (MSE).

square errors (*E*) and diversity (*D*) terms for external validations were determined (Figure 3).

Figure 3a shows test set *E* terms for 20 instances of NNEs using 1, 5, 10, 20, and 40 members. The test set *E* term is a measure of the likeness among external predictions and experimental activities. The *E* term for models using one member is, at the same time, the MSE for external predictions when single training and test sets are selected. Like in the analysis of *R* values, the *E* terms for single predictors vary dramatically. Several partitions produce test sets predicted with very low error, but when these sets are replaced, outcomes can be completely different. The gathering of members leads to a highly precise value of error. In fact, ensembles containing 40 predictors have similar test set *E* values around 0.296. This value is the result of the weighted contribution of all assembled predictors.

The diversity term (*D*) is a measure of the difference among the members of an ensemble. Figure 3b shows test set *D* terms for 20 instances of NNEs using 5, 10, 20, and 40 members. A high value of *D* contributes positively to the MSE value for unstable frames. In our case, the Bayesian regularization increased the stability of the predictors; therefore, no high values of *D* should be expected. In effect, *D* values are lower than 0.090 for all NNEs (Figure 3b). It is noteworthy that the gathering of members in the ensemble increases the mean value of *D* terms.

According to Krogh and Vedelsby analysis, MSE value for NNEs should be smaller than the averaged MSE of the component predictors (eq 12). Figure 3c shows test set MSEs for 20 instances of NNEs using 1, 5, 10, 20, and 40 members. The difference between this graphic and the previous one from Figure 3a is due to the *D* term contribution. In this sense, NNEs containing 40 members have similar test set MSE values around 0.232, which represents a decrease of a 22% in comparison to the mean *E* value. Since no difference was evidenced in the MSE values when there were 40 or 20 members assembled in the ensemble, 20 members seems to be enough to get a precise statistic of the validation. A plot of training and external predictions versus experimental activities employing 20-member NNEs is shown in Figure 1.

3D descriptors selected by GA were used to obtain SOMs, relating the descriptors with inhibitory activities in an unsupervised manner. In a self-organizing neural network, if two input data vectors are similar, they will be mapped into the same neuron or into very close neurons in the two-dimensional map. Therefore, either group in the map can be interpreted as a set of analogues defined by the six variables' vectorial space.

Figure 4 depicts a Kohonen SOM for the 98 CDK inhibitors. It is clearly visible that compounds are adequately distributed across the entire map: 72 out of a total of 144 neurons were occupied. As it is observed, compounds with a similar range of activities were grouped into neighboring areas. It is noteworthy that there is a kind of gradient from the less-active compounds at the lower-right zone to the most-active compounds at the upper-left zone (Figure 4b−h). As a consequence, this map can be used to carry out qualitative predictions. The position in the map would be able to assign an approximate range of activity for unknown compounds.

The external predictions of the 98 modeled cyclin-dependent kinase inhibitors are reported in Table 1. Compounds 24, 45, 66, 86, and 97 can be considered as test set outliers (residuals $> 2\ S_{CV}$) in our vectorial space. The reason for the wrong prediction of these compounds is related to inaccurate associations with similar molecules containing unequal activity. Some evidences can be derived from a closer inspection of the Kohonen SOM of Figure 4. Compounds 45, 66, and 97 were predicted as more active than they really are. This fact can be attributed to the high molecular similarity that these compounds share with more-active compounds, thus, avoiding a correct discrimination; it is hinted from the position they occupy in the map. In turn, the poorly active compound 45 [$\log(10^6/IC_{50}) = 3.62$],

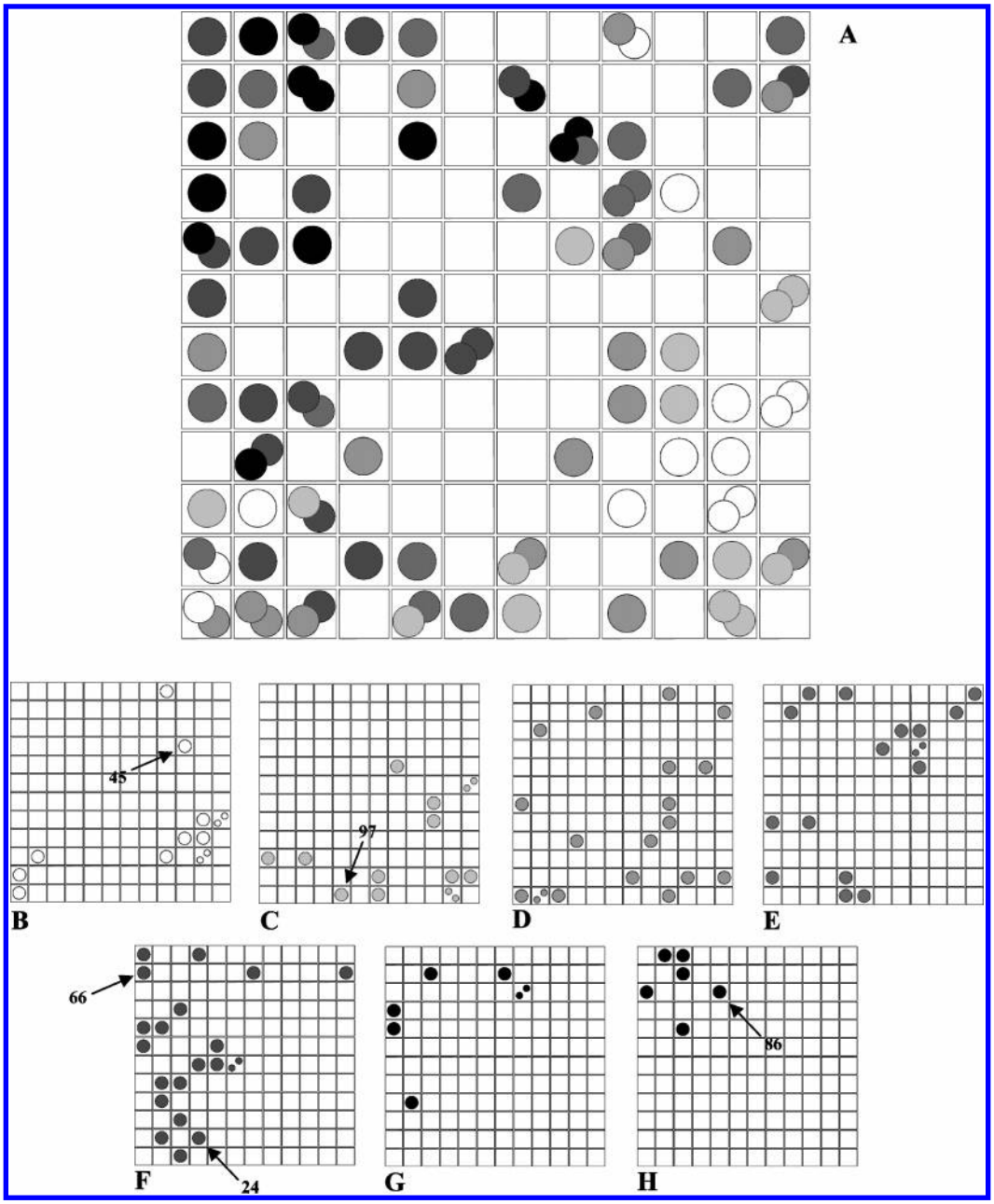


Figure 4. SOM for the data set using descriptors selected by genetic algorithm. Squares and circles denote neurons and compounds, respectively. (a) Map for the entire data set. Maps for compounds with $\log(10^6/IC_{50})$ values (b) from 3.60 to 4.00, (c) from 4.01 to 4.50, (d) from 4.51 to 4.96, (e) from 4.97 to 5.40, (f) from 5.41 to 5.80, (g) from 5.81 to 6.40, and (h) from 6.41 to 7.70. Test set outliers are pointed out.

which is the worst predicted in the whole data set, is located at a zone corresponding to more-active inhibitors [$\log(10^6/IC_{50}) > 5$] (Figure 4b). In addition, compound 66 is associated with the most-active inhibitors, although it is moderately active. On the other hand, compounds 24 and 86 are related to lesser-active inhibitors. This can be inferred by their surroundings in the SOMs: compound 86, which is one of the most-active inhibitors, is bordering the zone that contains compounds with a range of $\log(10^6/IC_{50})$ values around 6, while compound 24 is located in a complex zone where the majority of compounds present an inferior activity.

Despite the help that may offer an analysis of the SOMs, a simple comparison of the structures of the outliers with the associated compounds in the data set does not reveal why these compounds are poorly predicted. As a result of the complexity of the modeled activity and the structural diversity, any obvious explanation is not feasible.

CONCLUSIONS

Artificial neural networks are traditionally adequate for QSAR modeling. Their key strength is that, with the presence of hidden layers, ANNs are able to perform nonlinear mapping of the parameters to the corresponding biological activity implicitly.

We modeled the inhibitory activity against a cyclin-dependent kinase enzyme of a set of 1*H*-pyrazolo[3,4-*d*]-pyrimidine derivatives using back-propagated ANNs and Kohonen self-organizing maps. The molecular structures of these inhibitors were encoded in six 3D descriptors extracted

by genetic algorithm feature selection. A nonlinear dependence between CDK inhibitory activities and the 3D spatial information of the inhibitors was found.

The back-propagated model was able to explain about 87% of the data variance and, more importantly, it was quite stable to the inclusion/exclusion of compounds as measured by LOO cross-validation experiments ($Q^2 = 0.648$). Furthermore, we presented here a reasonable method to provide an additional validation process by means of neural network ensembles. For generating the NNE constituent predictors, we partitioned all of the data into several training and test sets. The assembled predictors aggregated their outputs to produce a single prediction. In this way, instead of predicting a sole, randomly selected external set, we predict the result of averaging several sets. The predictive power was measured by accounting for the averaged test set R values and test set MSE of 40 members' ensembles.

The results presented herein suggest a reliable method for validating QSAR models. The ensemble reached a middle value between the better and the worst individual ANN performances. In this sense, NNEs can be considered as an alternative in validation processes for getting more informative results than the classical Q^2 and the guileful external predictions in a sole validation set.

In connection with the above-reported relationship, we complement our analysis using Kohonen self-organizing maps. The inhibitors were adequately scattered in the map in accordance with their activity. In this sense, the SOMs can be taken as a useful tool for getting a qualitative analysis. The topology of the maps brings some information about the connection between compounds. The knowledge about the connections allows the visualization of the neighboring outliers.

ACKNOWLEDGMENT

We wish to thank Detlef Deymann and Arley Pérez for their helpful collaboration in the preparation of the manuscript.

REFERENCES AND NOTES

(1) Boyle, F. T.; Costello, G. F. Cancer therapy: A move to the molecular level. *Chem. Soc. Rev.* **1998**, *27*, 251−261.
(2) Foye, W. O. *Cancer Chemotherapeutic Agents*; American Chemical Society: Washington, DC, 1995.
(3) Harper, J. W.; Adams, P. D. Cyclin-Dependent Kinases. *Chem. Rev.* **2001**, *101*, 2511−2526.
(4) Sielecki, T. M.; Boylan, J. F.; Benfield, P. A.; Trainor, G. L. Cyclin-Dependent Kinase Inhibitors: Useful Targets in Cell Cycle Regulation. *J. Med. Chem.* **2000**, *43*, 1−18.
(5) Zetterberg, A.; Larsson, O.; Woman, K. G. What is the Restriction Point? *Curr. Opin. Cell Biol.* **1995**, *7*, 835−842.
(6) Senderowicz, A. M.; Headlee, D.; Stinson, S. F.; Lush, R. M.; Kalil, N.; Villalba, L.; Hill, K.; Steinberg, S. M.; Figg, W. D.; Tompkins, A.; Arbuck, S. G.; Sausville, E. A. Phase I Trial of Continuous Infusion Flavopiridol, a Novel Cyclin-Dependent Kinase Inhibitor, in Patients with Refractory Neoplasms. *J. Clin. Oncol.* **1998**, *16*, 2986−2999.
(7) Meijer, L.; Borgne, A.; Mulner, O.; Chong, J. P. J.; Blow, J. J.; Inagaki, N.; Inagaki, M.; Delcros, J. G.; Moulinoux, J. P. Biochemical and Cellular Effects of Roscovitine, a Potent and Selective Inhibitor of the Cyclin-Dependent Kinases Cdc2, Cdk2 and Cdk5. *Eur. J. Biochem.* **1997**, *243*, 527−536.
(8) Misra, R. N.; Xiao, H.; Kim, K. S.; Lu, S.; Han, W.; Barbosa, S. A.; Hunt, J. T.; Rawlins, D. B.; Shan, W.; Ahmed, S. J.; Qian, L.; Chen, B.; Zhao, R.; Bednarz, M. S.; Kellar, K. A.; Mulheron, J. G.; Batorsky, R.; Roongta, U.; Kamath, A.; Marathe, P.; Ranadive, S. A.; Sack, J. S.; Tokarski, J. S.; Pavletich, N. P.; Lee, F. Y. F.; Webster, K. R.; Kimball, S. D. *N*-(cycloalkylamino)acyl-2-aminothiazole inhibitors of cyclin-dependent kinase 2. N-[5-[[[5-(1,1-dimethylethyl)-2-oxazolyl]-methyl]thio]-2-thiazolyl]-4- piperidinecarboxamide (BMS-387032), a highly efficacious and selective antitumor agent. *J. Med. Chem.* **2004**, *47*, 1719−1728.
(9) Meijer, L.; Leclerc, S.; Leost M. Properties and potential applications of chemical inhibitors of Cyclin-Dependent Kinases. *Pharmacol. Ther.* **1999**, *82*, 279−284.
(10) Arris C. E., Boyle, F. T.; Calvert, A. H.; Curtin, N. J.; Endicott, J. A.; Garman, E. F.; Gibson, A. E.; Golding, B. T.; Grant, S.; Griffin, R. J.; Jewsbury, P.; Johnson, L. N.; Lawrie, A. M.; Newell, D. R.; Noble, M. E. M.; Sausville, E. A.; Schultz, R.; Yu, W. Identification of Novel Purine and Pyrimidine Cyclin-Dependent Kinase Inhibitors with Distinct Molecular Interactions and Tumor Cell Growth Inhibition Profiles. *J. Med. Chem.* **2000**, *43*, 2797−2804.
(11) Wang, S.; Meades, C.; Wood, G.; Osnowski, A.; Anderson, S.; Yuill, R.; Thomas, M.; Mezna, M.; Jackson, W.; Midgley, C.; Griffiths, G.; Fleming, I.; Green, S.; McNae, I.; Wu, S.-Y.; McInnes, C.; Zheleva, D.; Walkinshaw, M. D.; Fischer, P. M. 2-Anilino-4-(thiazol-5-yl)-pyrimidine CDK Inhibitors: Synthesis, SAR Analysis, X-ray Crystallography, and Biological Activity. *J. Med. Chem.* **2004**, *47*, 1662−1675.
(12) Pies, T.; Schaper, K.-J.; Leost, M.; Zaharevitz, D. W.; Gussio, R.; Meijer, L.; Kunick, C. CDK1−Inhibitory Activity of Paullones Depends on Electronic Properties of 9-Substituents. *Arch. Pharm. (Weinheim, Ger.)* **2004**, *337*, 486-492.
(13) Zupan, J.; Gasteiger, J. Neural networks: a new method for solving chemical problems or just a passing fase? *Anal. Chim. Acta* **1991**, *248*, 1−30.
(14) Fernández, M.; Caballero, J.; Helguera, A. M.; Castro, E. A.; González, M. P. Quantitative Structure−Activity Relationship to Predict Differential Inhibition of Aldose Reductase by Flavonoid Compounds. *Bioorg. Med. Chem.* **2005**, *13*, 3269−3277.
(15) Guha, R.; Jurs, P. C. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.
(16) Markwalder, J. A.; Arnone, M. R.; Benfield, P. A.; Boisclair, M.; Burton, C. R.; Chang, C.-H.; Cox, S. S.; Czerniak, P. M.; Dean, C. L.; Doleniak, D.; Grafstrom, R.; Harrison, B. A.; Kaltenbach III, R. F.; Nugiel, D. A.; Rossi, K. A.; Sherk, S. R.; Sisk, L. M.; Stouten, P.; Trainor, G. L.; Worland, P.; Seitz, S. P. Synthesis and Biological Evaluation of 1-Aryl-4,5-dihydro-1*H*-pyrazolo[3, 4-*d*]pyrimidin-4-one Inhibitors of Cyclin-Dependent Kinases. *J. Med. Chem.* **2004**, *47*, 5894−5911.
(17) Stewart, J. J. P. Optimization of parameters for semiempirical methods I-method. *J. Comput. Chem.* **1989**, *10*, 210−220.
(18) *MOPAC*, version 6; U.S. Air Force Academy: Colorado Springs, CO.
(19) Todeschini, R.; Consonni, V.; Pavan, M. *DRAGON*, version 2.1; Talete SRL: Milan, Italy.
(20) Kruszewski, J.; Krygowski, T. M. Harmonical Oscillator Approach to the Definition of Aromaticity. *Tetrahedron Lett.* **1972**, *36*, 3839−3842.
(21) Jug, K. A Bond Order Approach to Ring Current and Aromaticity. *J. Org. Chem.* **1983**, *48*, 1344−1348.
(22) Randic, M. Molecular Shape Profiles. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373−382.
(23) Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure−Activity Analysis; RSP−Wiley: Chichester, U. K., 1986.
(24) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from their Infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 151−164.
(25) Schuur, J.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.
(26) Todeschini, R.; Lansagni, M.; Marengo, E. New Molecular Descriptors for 2D and 3D Structures. Theory. *J. Chemom.* **1994**, *8*, 263−272.
(27) Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682−692.
(28) Holland, H. *Adaption in natural and artificial systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
(29) Cartwright, H. M. *Applications of artificial intelligence in chemistry*; Oxford University Press: Oxford, U. K., 1993.
(30) So, S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.
(31) *MATLAB*, version 7.0; The Mathworks Inc.: Natick, MA. http://www.mathworks.com.
(32) Hemmateenejad, B.; Safarpour, M. A.; Miri, R., Nesari, N. Toward an Optimal Procedure for PC-ANN Model Building: Prediction of the Carcinogenic Activity of a Large Set of Drugs. *J. Chem. Inf. Model.* **2005**, *45*, 190−199.

(33) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks applied to Structure−Activity Relationships. *J. Med. Chem.* **1990**, *33*, 905−908.
(34) (a) Mackay, D. J. C. Bayesian Interpolation. *Neural Comput.* **1992**, *4*, 415−447. (b) Mackay, D. J. C. A practical Bayesian Framework for Backprop Networks. *Neural Comput.* **1992**, *4*, 448−472.
(35) Lampinen, J.; Vehtari, A. Bayesian Approach for Neural Networks − Review and Case Studies. *Neural Networks* **2001**, *14*, 7−24.
(36) (a) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187. (b) Winkler, D. A.; Burden, F. R. Bayesian neural nets for modeling in drug discovery. *Biosilico* **2004**, *2*, 104−111.
(37) Foresee, F. D.; Hagan M. T. Gauss−Newton approximation to Bayesian learning. *Proceedings of the 1997 International Joint Conference on Neural Networks*; IEEE: Houston, 1997; pp 1930−1935.
(38) Hansen, L. K.; Salamon, P. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Machine Intell.* **1990**, *12*, 993−1001.
(39) Krogh, A.; Vedelsby, J. Neural network ensembles, cross-validation and active learning. In *Advances in Neural Information Processing Systems 7*; Tesauro, G., Touretzky, D., Lean, T., Eds.; MIT Press: Cambridge, MA, 1995; pp 231−238.
(40) Livingstone, D. J. *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*; Oxford University Press: Oxford, U. K., 1995.
(41) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191−193.
(42) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(43) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 357−369.
(44) Yan, A.; Gasteiger, J. Prediction of Aqueous Solubility of Organic Compounds Based on 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(45) Guha, R.; Serra, J. R.; Jurs, P. C. Generation of QSAR Sets with a Self-Organizing Map. *J. Mol. Graph. Model.* **2004**, *23*, 1−14.
(46) Agrafiotis, D. K.; Cedeño, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.
(47) Yan, A.; Gasteiger, J.; Krug, M.; Anzali, S. Linear and Nonlinear Functions on Modeling of Aqueous Solubility of Organic Compounds by Two Structure Representation Methods. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 75−87.
(48) de-Sousa, J. A.; Gasteiger, J. New Description of Molecular Chirality and Its Application to the Prediction of the Preferred Enantiomer in Stereoselective Reactions. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 369−375.
(49) Kohonen, T. Self-organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59−69.

CI050263I