Comparison of a Neural Net-Based QSAR Algorithm (PCANN) with Hologram- and Multiple Linear Regression-Based QSAR Approaches: Application to 1,4-Dihydropyridine-Based Calcium Channel Antagonists

Vellarkad N. Viswanadhan,*,[†] Geoffrey A. Mueller,[†] Subhash C. Basak,[‡] and John N. Weinstein*,[†]

Laboratory of Molecular Pharmacology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, and Center for Water and Environment, Natural Resources Research Institute, University of Minnesota, 5013 Miller Trunk Highway, Duluth, Minnesota 55811

Received June 30, 2000

A QSAR algorithm (PCANN) has been developed and applied to a set of calcium channel blockers which are of special interest because of their role in cardiac disease and also because many of them interact with P-glycoprotein, a membrane protein associated with multidrug resistance to anticancer agents. A database of 46 1,4-dihydropyridines with known Ca^{2+} channel binding affinities was employed for the present analysis. The QSAR algorithm can be summarized as follows: (1) a set of 90 graph theoretic and information theoretic descriptors representing various structural and topological characteristics was calculated for each of the 1,4-dihydropyridines and (2) principal component analysis (PCA) was used to compress these 90 into the eight best orthogonal composite descriptors for the database. These eight sufficed to explain 96% of the variance in the original descriptor set. (3) Two important empirical descriptors, the Leo-Hansch lipophilic constant and the Hammet electronic parameter, were added to the list of eight. (4) The 10 resulting descriptors were used as inputs to a back-propagation neural network whose output was the predicted binding affinity. (5) The predictive ability of the network was assessed by cross-validation. A comparison of the present approach with two other QSAR approaches (multiple linear regression using the same variables and a Hologram QSAR model) is made and shows that the PCANN approach can yield better predictions, once the right network configuration is identified. The present approach (PCANN) may prove useful for rapid assessment of the potential for biological activity when dealing with large chemical libraries.

1. INTRODUCTION

Over the last several years, we have used neural networks and other multi-variate methods of analysis to analyze the relationship between function and mechanism of action in the large, diverse set of drug molecules tested in the National Cancer Institute's cancer drug discovery program.[1−5] The next logical step was to integrate structure−function and structure−activity relationships into the analysis as well. It has also become apparent that rapid advances in recent years of high throughput synthesis and screening methodologies has meant that rapid QSAR model construction and analysis are crucial for success of these efforts. As a start in that direction, we have begun with a test case based on a set of homologous 1,4-dihydropyridine calcium (Ca^{2+}) channel blockers. The 1,4-dihydropyridines are well-known as calcium channel blockers and also as inhibitors of P-glycoprotein, a membrane transport protein coded by *MDR-1* and associated with multidrug resistance to anticancer agents.[6,7] P-glycoprotein has a 1,4-dihydropyridine-selective drug acceptor site that appears to be allosterically coupled to other binding sites.[8] Dihydropyridines may, therefore, be useful as multidrug resistance reversal agents. Nifedipine, for example, has been found to enhance the antitumor activity of cisplatin[9] and other anticancer agents.[10]

The dihydropyridines examined here have been the subject of several SAR, QSAR, and 3D-QSAR studies.[11−15] The data set for the present analysis included 2,6-dimethyl-3,5-dicarbomethoxy-4-phenyl-1,4-dihydropyridine and its ortho-, meta-, and/or para-substituted derivatives. An earlier QSAR analysis of this data set by Coburn and co-workers[14] used two well-known empirical descriptors, the Leo-Hansch lipophilicity term (π) and the Hammet electronic parameter (σ), as well as three other sterimol parameters. Since our aim was to build toward the analysis of a bigger and more diverse set of compounds, we employed a much larger set of molecular properties, which included a set of 90 topological indices from graph theoretic and information theoretic analyses.[16−19] Topological indices have been found useful in classifying chemical structures[17,20,21] and in predicting physicochemical and biological properties.[22,23]

The most familiar standard approaches to QSAR[24] are based on multiple linear regression (MLR) and partial least squares (PLS) regression.[25,26] However, these approaches can capture only linear relationships between molecular characteristics and structural or functional features to be predicted. In contrast, neural networks[27,28] are capable of recognizing highly nonlinear relationships; hence, they provide an interesting approach to QSAR,[29−32] quantitative structure−

* Corresponding authors: Vellarkad N. Viswanadhan (current address: Amgen, Inc., 1 Amgen Center Drive, M/S 14-2-B, Thousand Oaks, CA 91320) or John N. Weinstein at the National Cancer Institute.

[†] National Institutes of Health.

[‡] University of Minnesota.

10.1021/ci000072+ CCC: $20.00
Published on Web 04/19/2001

property relationships (QSPR).[33,34] In this study, we first subjected the matrix of theoretical descriptors to a principal component analysis (PCA), thereby reducing the number of variables. We then added the empirical descriptors and used the combination as inputs to the network, thereby developing the PCANN predictive models. Previously PCA and NN approaches were used separately or in some instances, in combination,[35] though not specifically for QSAR. PCA, it should be noted, uses information only from the internal variations in the input matrix; unlike PLS, it does not take account of the dependent variable (i.e., the measured binding constant). Using the same variables, an MLR model was also developed for comparison. Last, a molecular fragment-based QSAR approach known as HQSAR (Hologram-QSAR) was also used to develop predictive models, and all three types of models were compared in terms of ease of use and overall predictive power.

2. METHODS

2.1. Overview of the QSAR Algorithm (PCANN). The PCANN algorithm developed and used in this study can be summarized as follows:

(1) Ninety topological indices (TIs) based on information theoretic and graph theoretic analysis[36] were calculated for molecules in the database using the program POLLY.[18]

(2) Principal component analysis (PCA) was then used to assess the intrinsic dimensionality of the problem and extract a linear combination of the descriptors that explain most of the variance in the original data. In the PCA presented here, we started with the 90×90 correlation matrix[37] and used the dominant principal components (PCs) as variables for each molecule.

(3) Two well-known substituent constants, the Leo-Hansch lipophilic term (π) [38] and the Hammet electronic parameter (σ),[39] were added to the variable list from step (2). Properties dependent on conformation, stereochemical configuration, and charge distribution (such as dipole moment and hydrophobic moment) could also be added as appropriate.

(4) A feed-forward, back-error propagation neural network[40] was constructed to model the structure−activity relationships. The input vector was the set of descriptors for each molecule in the series, as generated by the previous steps. The network was configured with one hidden layer of processing elements (PEs). The presence of a hidden layer makes it possible to classify categories (or outputs) that are not linearly separable. The network was trained to reproduce the binding affinities by repetitive presentation of the set of input vectors in random order within each presentation of the entire set.

(5) Cross-validation was performed by dividing the input dataset into five distinct training and test subsets such that each training set covered all of the substituent positions and represented the structural diversity in the original dataset.

Figure 1 shows the skeletal structure 2,6-dimethyl-3,5-dicarbomethoxy-4-phenyl-1,4-dihydropyridine. This compound and the 45 derivatives listed in Table 1 constituted the database for analysis. Table 1 lists values of log(1/EC_{50}), a pharmacological measure of the effect of a calcium channel antagonist on the tonic contractile response of longitudinal muscle strips from guinea pig ileum. These data and values for two empirical parameters, the lipophilic substituent constant (π) and the Hammet electronic constant (σ), for each compound in the dataset were calculated in an earlier study.[14] In addition, 90 theoretical descriptors based on information-theoretic and graph theoretic analyses were determined for each compound, as described below.

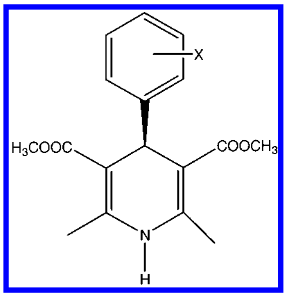

Figure 1. The skeletal 1,4-dihydropyridine structure with substitutions/changes indicated by X on the phenyl ring. These substitutions are listed in Table 1.

2.2. Graph Theoretic and Information Theoretic Indices. All of the TIs used here have been well-documented[41] and hence are not discussed in detail here. They quantitate two types of molecular properties: connectivity and complexity. Molecular connectivity is quantitated using chemical graph theory,[42] and molecular complexity is quantitated using chemical information theory.[43] The TIs used in our analysis were derived from the adjacency matrix and distance matrix of a chemical graph, calculated using the computer program POLLY.[18] Input to the program was the chemical structure encoded in SMILES[44] representation. Table 2 lists types of indices included in the present analysis. Among them were the Weiner index,[45] several other connectivity indices[21] and molecular complexity indices from chemical information theory.[43] These indices have been widely used in chemical classification,[17,20,21] QSAR,[23] and QSPR studies.[22,46]

The Weiner index (W), the information-theoretic index on graph distance ($I^W{}_D$), the mean information index ($I^W{}_D$), and the path length parameters of order h (P_h) were derived from the topological distance matrix, D(G), of the hydrogen-suppressed chemical graph. D(G) is a symmetric $n \times n$ matrix, where n is the number of non-hydrogen atoms in the molecule. Each element of D(G) is the topological distance (smallest number of intervening bonds in the graph, G) between two vertices (say, v_i and v_j). P_h is the number of paths of length h in the graph. In this study, indices covering paths of length $h = 0, 1 ...., 6$ were included.[36] To calculate the information indices, a set A of n elements was derived from the molecular graph G, which depends on specific structural characteristics. This set was partitioned into disjoint subsets A_i of order n_i ($i = 1,2,...h$: $\sum n_i = n$), based on an equivalence relation defined on A. A probability distribution was then assigned to the equivalence classes. This distribution gives the probability that a randomly selected element of A (A_i) will occur in the ith subset. Information indices were then computed from the probability distribution. The indices used here included those derived from the distance matrices, measures of graph complexity, structural information measures, and complementarity information measures.[36]

2.3. Principal Component Analysis. The descriptor data for the set of n molecules defines a 90-dimensional parameter space (R^{90}). Each compound corresponds to a point in that R.[90] The method of principal component analysis (PCA, or Karhunen-Loeve transformation)[47] reduces the dimensionality

Table 1. Results of MLR, Neural Network, and HQSAR Prediction Models for Compounds Identified by the Substituent "X" in Figure 1[a]

	substituent X	expt. log(1/EC_{50})	predicted by MLR	PCANN (different PEs) 2	3	4	8	10	predicted by HQSAR	test set
1	3-Br	8.89	6.80	6.86	6.85	6.72	7.08	7.14	7.09	3
2	2-CF_3	8.82	6.96	7.21	6.88	7.16	7.13	7.13	7.66	1
3	2-Cl	8.66	8.74	8.40	8.37	8.29	8.32	8.39	7.67	2
4	3-NO2	8.40	7.14	7.39	7.51	7.46	7.66	7.60	7.35	2
5	2-$CH{=}CH_2$	8.35	8.48	8.05	7.91	8.00	7.62	7.65	7.83	3
6	2-NO_2	8.29	7.71	7.97	8.60	8.24	8.71	7.81	8.62	4
7	2-CH_3	8.22	7.79	7.88	7.59	7.50	7.26	7.18	7.70	5
8	2-Et	8.19	7.79	7.59	7.59	7.69	7.92	7.86	6.88	1
9	2-Br	8.12	8.59	8.57	8.52	8.42	8.45	8.51	7.71	2
10	2-CN	7.80	7.79	8.46	8.23	8.00	8.36	8.34	7.80	3
11	3-Cl	7.80	7.80	7.62	8.10	7.73	8.28	8.24	7.17	1
12	3-F	7.68	7.44	7.44	7.09	7.54	7.10	7.69	7.18	4
13	H*	7.55	******	******	******	******	******	******	7.30	*
14	3-CN	7.46	7.65	7.62	7.58	7.76	7.54	7.61	7.30	5
15	3-I	7.38	6.38	6.71	6.60	6.38	6.71	6.81	7.20	3
16	2-F	7.37	8.15	8.43	7.89	8.51	7.90	8.29	7.76	4
17	2-I	7.33	7.45	7.59	7.34	7.92	8.01	8.01	7.21	3
18	2-OCH_3	7.24	7.01	6.79	6.71	7.34	6.60	6.72	7.13	5
19	3-CF_3	7.13	8.39	8.45	8.62	8.48	8.57	8.59	5.36	2
20	3-CH_3	6.96	7.52	7.43	7.58	7.55	7.56	7.54	7.23	1
21	2-OEt	6.96	7.33	7.34	7.16	7.39	7.41	7.40	7.06	1
22	3-OCH_3	6.72	6.44	6.46	6.18	6.27	5.86	6.57	6.67	4
23	3-N Me_2	6.05	4.84	5.31	4.55	5.38	4.33	5.11	5.86	5
24	3-OH	6.00	6.77	6.49	6.62	6.50	6.23	6.14	7.31	3
25	3-NH_2	5.70	5.35	5.85	5.14	5.59	4.98	4.77	7.33	2
26	3-OAc	5.22	6.26	6.52	6.71	6.05	6.29	6.43	6.73	1
27	3-O−COPh	5.20	10.35	7.49	8.99	8.83	7.23	8.80	5.31	4
28	2-NH_2	4.40	6.60	6.29	5.89	6.16	5.45	5.25	7.95	2
29	3-N^+Me3	4.30	6.06	4.54	4.61	4.26	4.65	4.31	6.66	5
30	4-F	6.89	6.17	5.54	4.96	5.16	6.18	5.16	5.36	4
31	4-Br	5.40	5.80	5.31	5.37	5.66	4.93	5.68	5.58	5
32	4-I	4.64	5.95	5.41	5.70	5.60	5.66	5.67	5.69	1
33	4-NO_2	5.50	5.63	6.03	6.19	6.14	6.34	6.24	5.81	2
34	4-N Me_2	4.00	4.51	3.43	3.28	3.44	2.94	2.96	3.98	3
35	4-CN	5.46	5.92	4.97	5.42	5.18	5.90	4.16	5.85	4
36	4-Cl	5.09	6.31	6.74	6.40	6.60	5.80	6.76	5.62	5
37	2,6-Cl_2	8.72	6.68	5.09	5.70	7.43	6.05	6.98	8.58	5
38	F_5	8.36	7.29	9.40	9.46	9.11	8.53	9.06	6.36	4
39	2-F,6-Cl	8.12	8.37	7.83	7.51	7.64	6.97	7.03	8.73	3
40	2,3-Cl_2	7.72	9.74	8.50	8.49	8.40	8.60	8.57	8.23	2
41	2-Cl,5-NO_2	7.52	7.51	7.72	7.07	7.72	6.86	6.93	7.24	1
42	3,5-Cl_2	7.03	5.95	4.79	6.01	6.10	6.49	4.32	7.44	5
43	2-OH,5-NO_2	7.00	6.03	7.24	7.86	7.16	8.18	6.52	7.47	4
44	2,5-Me_2	7.00	6.97	7.05	6.92	6.80	6.67	6.81	7.08	3
45	2,4-Cl_2	6.40	8.59	8.40	8.41	8.37	8.42	8.41	6.22	2
46	2,4,5-$(OCH_3)_3$	3.00	4.68	5.22	5.45	4.47	4.53	4.45	3.88	1

[a] Log(1/EC_{50}) is the 50% effective concentration for blocking Ca^{2+} channel. The predicted values were obtained from multiple linear regression (MLR), neural network models with 2, 3, 4, 8, or 10 processing elements in the hidden layer of the back-propagation network and the HQSAR model. The last column indicates the test set in which each compound fell (for MLR and neural net models). The HQSAR predictions are based on leave-one-out cross-validation.

by embedding all of the points (compounds) in a subspace dimensionally less than 90 such that a desired degree of variance in the original data is captured in the subspace. Each principal component (PC) is a linear combination of the original variables, with coefficients given by the eigenvectors of the covariance matrix (in this case equivalent to the correlation matrix). Given the normalization of the variables, the first PC is the axis that minimizes the sum of squared Euclidean distances from the points to that axis. Subsequent PCs similarly minimize the residual variance.

In doing PCA, scaling choices must be made because PCs are scale-dependent. To control this dependence, the most commonly used convention is to rescale the variables so that each variable has mean zero and standard deviation one. In some cases, outliers will have large effects on best fitting planes when Euclidean metrics are used. For distributions that are positive and highly skewed by large values, a logarithmic transformation can be useful in reducing the importance of outliers. For the data in this investigation, it was desirable to transform the indices by taking log(index + 1) and then standardizing to mean zero and variance one. For reduction of dimensionality, we retained the PCs with eigenvalues greater than unity.[48]

2.4. Neural Network Modeling. Feed-forward, back-propagation-of-error networks were developed using the NeuralWare Professional II package.[49] The training procedure used was similar to that of Rumelhart et al.[40] Network weights ($W_{ji}(s)$) for a processing element "j" receiving output from PE "i" in the layer "s" were initially assigned random values between −0.1 and +0.1. We chose the

Table 2. Brief Description of Types of Graph Theoretic and Information Theoretic Variables Used in the Principal Component Analysis

W	Weiner index; half-sum of the off-diagonal elements of the distance matrix of a graph
$I_D{}^W$	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$I_D{}^W$	mean information index for the magnitude of distance
IC_r	mean information content or complexity of a graph based on the *r*th ($r = 0,1,2,.0.6$) order neighborhood vertices of a graph
SIC_r	structural information content of a graph based on*r*th ($r = 1,2,..0.6$) order neighborhood of vertices
CIC_r	complementary information content of a graph g calculated from *r*th ($r = 1,2,..0.6$) order neighborhood of vertices
${}^h\chi$	path terms of hth order ($h = 1,2,..0.6$)
${}^h\chi_C$	chain or cycle terms of hth order ($h = 3,..0.6$)
${}^h\chi_{PC}$	path-cluster terms of hth order ($h = 1,2,..0.6$)
${}^h\chi_{CH}$	chain or cycle terms of hth order ($h = 3,..0.6$) calculated from values
${}^h\chi^b{}_C$	bonding connectivity type cluster terms of hth order ($h = 1,2,..0.6$)
${}^h\chi^b{}_{PC}$	bonding connectivity type path-cluster terms of hth order ($h = 4,..0.6$)
${}^h\chi^b{}_{CH}$	bonding connectivity type chain or cycle terms of hth order ($h = 3,..0.6$)
$h\chi^v$	valence connectivity type chain or cycle terms of hth order ($h = 3,..0.6$)
${}^h\chi\upsilon_C$	valence connectivity type cluster terms of hth order ($h = 1,2,..0.6$)
${}^h\chi^b\mathrm{PC}$	valence connectivity type path-cluster terms of hth order ($h = 4,..0.6$)
${}^h\chi^b\mathrm{CH}$	valence connectivity type chain or cycle terms of hth order ($h = 3,..0.6$)
O	order or neighborhood when IC_r reaches its maximum value
P_h	number of paths of length h ($h = 0,1,...,10$) in the hydrogen deleted graph

hyperbolic tangent as the transfer function that generates the output of a neuron from the weighted sum of inputs from the preceding layer of PEs. Consecutive layers were fully interconnected; there were no connections within a layer or between the input and the output. A bias unit with a constant activation of unity was connected to each unit in the hidden and output layers.

During the learning process, each compound in the training set was iteratively presented to the network. That is, the compound's vector of 10 descriptors (in normalized form) was fed to the input PEs, and the network's output was compared with the experimental "target" value. During one "epoch", all compounds in the training set were presented, and weights in the network were then adjusted on the basis of the discrepancy between network output for log(affinity) and the experimental value. The training set was presented in a different random order during each epoch to avoid cyclical behavior and the local minima that could result. Weight adjustment proceeded on the basis of a gradient descent in the rms error between output values (predicted binding affinities) and target values (experimental affinities).

For the gradient descent algorithm, a very low learning preinitiation rate was used to ensure that the error surface was sufficiently linear locally. This ensured slow convergence of the network and increased the likelihood of finding a good minimum. Convergence properties of the network were enhanced by using a "momentum" term (m) proportional to the previous weight change. The weights were thus modified according to the expression

$$\Delta w_{ji}(s) = a_* e_{j*} x_i(s) + m_* \Delta w_{ji}(s)'$$

averaged over the epoch. In this equation, *a* is the learning coefficient and the Δw's are weight changes for the present ($\Delta w_{ji}(s)$) and previous iteration ($\Delta w_{ji}(s)'$), respectively. We used learning rates such that the hidden layer tended to train faster than the output as is generally desirable. The learning algorithm was not tuned for the data set.

Input to the network consisted of (i) the first eight principal components derived from principal component analysis of 90 graph theoretic and information theoretic variables, as described in the previous section and (ii) two other variables, the lipophilic substituent constant (π) and the Hammet electronic parameter (σ), which appear to be independently important.

Several network configurations were tested, each with a different number of hidden layer elements. This analysis was confined to a single hidden layer because networks with more than one hidden layer would be harder to train and almost certainly of no additional value for this class of problems. The size of the hidden layer controls the number of potential degrees of freedom used in the model. Too large a hidden layer could reduce the residual errors with parameters specific to the training set only, thus reducing the predictive ability of the model. That is, a large network tends to learn particular features of the training set and to lose its capacity for generalization because each weight is a potential free variable. However, the learning process tends to bring into play (i.e., as substantially different from zero) only the approximate number of weights necessary to explain the variance. Hence, the effective number of free parameters will generally be lower than the number of weights. To ensure that over-training of the networks did not occur, for each configuration correlation coefficients and rms deviations with experimental values were monitored for training and test sets. The training was stopped after no improvement in rms deviation was found for the test sets.

2.5. Cross-Validation for PCANN and MLR Models. For the purpose of testing various models, five different pairs of training and test sets were constructed from the data of Table 1. The union of each training set and corresponding test set equaled the original database, and each of the molecules was present once in one of the test sets, except for the unsubstituted original compound (entry 13 in Table 1), which was present in all training sets. At least one entry corresponding to each substituent position was placed in each training set (i.e., the database was stratified by substituent position). The last column of Table 1 identifies the test set to which each compound belonged.

2.6. Molecular Holograms and the HQSAR Model. The molecular holograms[50] is a special form of 2D/3D fingerprints based on counting the number of times each unique fragment occurs by "binning" fragments of different sizes and composition. Dissimilar fragments may sometimes fall into the same bin because of the restriction on hologram size (always predefined). Nevertheless, holograms are useful because they contain attributes useful for QSAR or QSPR studies, and they can be generated for large chemical libraries in an automated fashion. For the present work, HQSAR module available in Sybyl (version 6.6) was used. A Sybyl molecular spreadsheet for the 46 molecules in Table 1 was constructed using Sybyl tools,[51] which was used for developing HQSAR models. Several hologram models with various hologram lengths and fragment sizes were generated and tested by leave-one-out cross-validation and the HQSAR model that gave the best cross-validated results (in terms of cross-validated standard error) was chosen.

2.7. Assessment of the QSAR Models. Different QSAR models (based on MLR, PCANN, and HQSAR) used in the analyses were assessed by Pearson correlation coefficient (R), RMS deviation (RMS), maximum and minimum deviations (max. dev. and min. dev.) for each set, and "predictive r^2". These parameters were calculated for each set of cross-validated predictions.

Predictive r^2 measures the quality of predictions relative to a simple "no model" guess, given by

$$\text{predictive } r^2 = (\text{SD-“press”})/\text{SD}$$

where SD is the sum of squared deviations of each measured $\log(1/IC_{50})$ value from its mean and "press" is the predictive sum of squared differences (the sum of squared differences between actual and predicted values). Negative values for predictive r^2 indicate that $\log(1/IC_{50})$ is better estimated by "mean of values" than by the model under consideration.

3. RESULTS AND DISCUSSION

The compounds represented in this study share the same molecular skeleton, 2,6-dimethyl-3,5-dicarbomethoxy-4-phenyl-1,4-dihydropyridine. Earlier QSAR analysis[14] clearly indicated the need for inclusion of lipophilic, electronic, and steric terms in the QSAR equation. Specifically, the Hammet constant for the substituent at the *meta* position has been shown to be important. In the present study, we included this parameter and the lipophilic substituent constant explicitly to model electrostatic and hydrophobic interactions; these characteristics are unlikely to be well-represented in the information-theoretic and graph-theoretic descriptors. It was not necessary to represent steric effects separately as they are represented by topological indices which quantitate the size and shape of a molecule. Topological indices encode structural characteristics related to connectivity and complexity that could be important for binding,[42,50] and they can be readily calculated for very large data sets.

Table 3 summarizes the results of principal component analysis of the 90 graph theoretic and information theoretic variables calculated for the 46 compounds shown in Table 1. Listed are the eigenvalues (which were used as input variables to the neural network) and the cumulative explained variance of the first eight PCs. The eight PCs used here explained 96% of the variance in the set. Thus, these eight PCs sufficed to represent most of the information in these topological indices.

Table 3. Results of Principal Component Analysis[a]

		variance explained				variance explained	
PC	eigenvalue	%age of	cum	PC	eigenvalue	%age of	cum
1	35.2	43.4	43.4	5	5.8	7.1	87.5
2	13.7	17.0	60.4	6	3.9	4.8	92.4
3	9.1	11.2	71.6	7	2.0	2.5	94.9
4	7.1	8.8	80.4	8	1.0	1.3	96.1

[a] For each component (PC), the corresponding eigenvalue, the percentage of variance explained, and the cumulative percentage up to that component are shown.

Neural networks described here were constructed using NeuralWare Professional II on a desktop personal computer, as described in Methods. Predicted binding affinities (expressed as the logarithm of effective concentration of the drug candidate required for 50% inhibition) from different network configurations were stored. The complete set of predictions (for all compounds minus the skeletal standard) was generated by combining the stored predictions from test data for each of the five networks. To avoid over-training the networks, the length of training (number of epochs used in training the datasets) was chosen to maximize overall test set rms deviations. The same number of iterations were used for all training sets, so only one degree of freedom was lost by this training. In Table 1, different columns correspond to the observed binding affinity, the binding affinity predicted by MLR, by neural networks with different numbers of PEs, and by the HQSAR method. The HQSAR model reported here was obtained by examining several different HQSAR models with different hologram lengths (lengths of 61, 71, 151, 257, 353) and different sets of fragment lengths (1 to 7 atoms, 1 to 4 atoms). The multi-HQSAR option in Sybyl[50,51] was used in examining all these different models. Of 14 different models examined, the best cross-validated model gave a cross-validated q^2 of 0.489. This model had a hologram length of 257, used molecular fragments of lengths 4 to 7 and used four PLS components.

The PCANN and MLR models used a cross-validation scheme outlined in section 2.5 above. For HQSAR models, cross-validation was performed by the more usual leave-one-out formalism. Leaving several molecules out at once for these datasets resulted in much poorer cross-validation statistics for HQSAR models. In the HQSAR method, molecules are represented as combinations of specific fragment types (holograms), whereas the PCANN method uses overall molecular properties. Hence, the HQSAR method needs a greater number of distinct molecules (holograms) for training than the PCANN method. On the other hand, the HQSAR method can have some advantages with outliers (i.e., molecules/substituents distinctly different from the rest). Such molecules are usually are not very active (Otherwise, more similar molecules would have been synthesized and tested, and they would not be distinct.). In these unusual cases, distinct molecular fragments would be assigned to rarely occupied bins of the holograms. They are typically assigned smaller coeffcients, and hence they will generally be predicted to be less active. As the PCANN/MLR methods developed here rely on overall molecular properties, the outliers can lead to extreme predictions (highly active or totally inactive). **27** (Table 1), a compound with low activity, presents such an example,

Table 4. Statistical Results for the Calculated (Predicted) Values for 45 Compounds Predicted by Various Models

method	correl.	rms	max. dev.	min. dev.	pred. r^2	no. points
MLR	0.55	1.30	5.15	0.00	0.19	45
PCANN (2PE)	0.65	1.15	3.63	0.05	0.36	45
PCANN (3PE)	0.61	1.22	3.79	0.01	0.28	45
PCANN (4PE)	0.71	1.05	3.63	0.04	0.47	45
PCANN (8PE)	0.73	1.02	2.67	0.08	0.50	45
PCANN (10PE)	0.68	1.16	3.60	0.01	0.35	45
HQSAR_BEST	0.70	1.02	3.55	0.00	0.49	46

for which notable disagreement between observed and calculated values occurs for both PCANN and MLR methods. The neural network prediction is better than that of MLR. **27** has the only cyclic substituent (−OCOPhe) in the entire data set and also has the largest number of heavy atoms of any substituent. Consequently, neither PCANN nor MLR was able to predict its activity. For the reasons mentioned earlier, a much better agreement is observed with the HQSAR model in this case. Based on the previous work,[14] it was clear that predictive QSAR models for the dataset required addition of the empirical descriptors σ and π to other theoretical or structural types of descriptors. Strong correlations (up to 0.95 for the correlation coefficient, R) have been reported in earlier QSAR work[14] for this data set. However, in contrast to the present work, those were not cross-validated, and hence no direct comparison of the results is feasible.

Table 4 lists cross-validated statistical results of the present analysis; it represents "test" correlations (cross-validated) for each molecule. Table 4 indicates that neural net models performed better than MLR and that the best PCANN model (with eight hidden units) is marginally better than the best HQSAR model. The results were relatively insensitive to the number of hidden layer PEs over a wide range. In the literature, one finds other examples in which network performance is relatively insensitive to the number of hidden layer PEs. Studies by Andrea and Kalayeh,[30] for example, showed an instance in which the reduction in test-set variance between 3 and 8 hidden layer PEs was small. In these cases, it is obviously preferable to employ networks with fewer adjustable parameters, that is, with smaller numbers of hidden PEs. Since it is not clear how many independent components are represented in the set of weights (some of which remain approximately equal to zero in the larger networks), it is uncertain what penalty to impose in the formal sense for addition of hidden layer PEs and consequent loss in the degrees of freedom. The "predictive r^2" measure introduced by Cramer et al.[25] is a more stringent measure of the predictive ability of each model. As shown in Table 4, even for the best models obtained here, predictive r^2 values do not exceed 0.5. Greater predictive ability could perhaps be obtained with 3D QSAR/receptor site models in this case.

An earlier PCA of the same set of TIs for a larger database of 3692 chemical structures[52] yielded 10 eigenvalues greater than 1.0. The corresponding PCs explained 93% of the variance in the original data. Analysis of these PCs revealed that the first component largely represented shape and size, the second represented symmetry, the third represented degree of branching, and the fourth represented cyclicity. The four PCs derived for this dataset also have similar characteristics and represent the same four classes of physical properties. Chemical graph theory and information theory provide a way of representing such molecular properties in terms of TIs that can be the used in QSAR. The extent of cross-validated correlations seen in this study points to the role played by these properties in ligand binding and, consequently, their potential utility in ligand design.

4. CONCLUSION

We present here a QSAR algorithm, PCANN, which combines principal component analysis (PCA) with neural networks (NN) pattern recognition. The algorithm was applied to a set of 1,4-dihydropyridines that are well-known as calcium channel blockers and also as inhibitors of P-glycoprotein transport function. We then compared the results with those from two other models. In a cross-validated study, the neural network models were shown to be better than the corresponding multiple linear regression in predicting binding affinities from a set of molecular descriptors that included σ, π and eight principal components obtained from 90 TIs. The PCANN models were also shown to be better (in terms of cross-validated statistics) than the hologram QSAR (HQSAR) models for this dataset. Setting up HQSAR models, on the other hand, is more straightforward and easier. Nonetheless, the results with PCANN reflect the utility of TIs that encode such topological properties as shape, symmetry, degree of branching, and cyclicity. Some of these TIs have been applied productively in the context of combinatorial chemistry to characterize large chemical libraries for the identification and optimization of lead compounds.[53−55] Cerius2 combichem tools[56] use both TIs and other empirical descriptors such as lipophilicity for QSAR and library design applications. Validation of these metrics is a subject of ongoing research.[57] The present dataset is clearly too limited for any general conclusions to be drawn with regard to the utility of the approaches presented here. Nevertheless, this study supports the idea that these nonempirical descriptors usefully complement the more widely known empirical descriptors of lipophilic and electronic properties for use in the design of focused libraries or "computational screening" of large databases prior to biochemical testing.

ACKNOWLEDGMENT

A shorter, electronic version of the article appeared in the journal "Network Science" edited by Allen B. Richon. The link to the article is http://www.netsci.org/Science/Compchem/feature07.html.

REFERENCES AND NOTES

(1) Weinstein, J. N.; Kohn, K. W.; Driscoll, J. S.; Grever, M. R.; Viswanadhan, V. N.; Rubinstein, L. V.; Paull, K. D. *Science* **1992**, *258*, 447.

(2) Weinstein, J. N.; Rubinstein, L. V.; Koutsoukos, A. D.; Kohn, K. W.; Grever, M. R.; Monks, A.; Scudiero, D. A.; Welch, L.; Chiausa, A. J.; Fojo, A. T.; Viswanadhan, V. N.; Paull, K. D. *World Congress Neural Networks* **1993**, *1*, 111.

(3) Weinstein, J. N.; Myers, T. G.; O'connor, P. M.; Friend, S. H.; Fornace, Jr. A. J.; Kohn, K. W.; Fojo, T.; Bates, S. E.; Rubinstein, L. V.; Anderson, N. L.; Buolamwini, J.; van Osdol, W. W.; Monks, A. P.; Scuderio, D. A.; Sausville, E. A.; Zaharevitz, D. W.; Bunow, B.; Viswanadhan, V. N.; Johnson, G. S.; Wittes, R. E.; and Paull, K. D. *Science* **1997**, *275*, 343.

(4) van Osdol, W. W.; Myers, T. G.; Paull, K. D.; Kohn, K. W.; Weinstein, J. N. *J. Natl. Cancer Inst.* **1994**, *86*, 1853.

(5) Koutsoukos, A. D.; Rubinstein, L. V.; Faraggi, D.; Kalyandrug, S.; Weinstein, J. N.; Paull, K. D.; Kohn, K. W.; Simon, R. M. *Statistics Medicine* **1994**, *13*, 719.

(6) Kertner, N.; Riordan, J. R.; Ling, V. *Science* **1983**, *221*, 1285.

(7) Gottasman, M. M.; Liryeyna, C. A.; Germann, P. V.; Pastan, I. *Annu. Rev. Genet.* **1995**, *29*, 607.

(8) Ferry, D. R.; Russell, M. A.; Cullen, M. H. *Biochem. Biophys. Res. Commun.* **1992**, *188*, 440.

(9) Onoda, J. M.; Nelson, K. K.; Taylor, J. D.; Honn, K. V. *Cancer Res.* **1989**, *49*, 2844.

(10) Fine, R. L.; Koizumi, S.; Curt, G. A.; Chabner, B. A. *J. Clin. Oncol.* **1987**, *5*, 489.

(11) Mahmoudian, M.; Richards, G. W. *J. Pharm. Pharmacol.* **1986**, *38*, 372.

(12) Norrington, F. E.; Hyde, R. M.; Williams, S. G.; Wooton, R. *J. Med. Chem.* **1975**, *18*, 604.

(13) Rovnyak, G.; Anderson, N.; Gougoutas, J.; Hedberg, A.; Kimball, S. D.; Malley, M.; Moreland, S.; Porubcan, M.; Pudzianowski, A. *J. Med. Chem.* **1991**, *34*, 2521.

(14) Coburn, R. A.; Weirzba, M.; Suto, M. J.; Solo, A. J.; Triggle, A. M.; Triggle, D. J. *J. Med. Chem.* **1988**, *31*, 2103.

(15) (a) Belvisi, L.; Brossa, S.; Salimbeni, A.; Scolastico, C.; Todeschini, R. *J. Comput.-Aided Mol. Design* **1991**, *5*, 571. (b) Costa, M. C. A.; Gaudio, A. S.; Takahata, Y. *J. Mol. Struct. (THEOCHEM)* **1997**, *394*, 291.

(16) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Modeling* **1987**, *8*, 300.

(17) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Appl. Math.* **1988**, *19*, 17.

(18) Basak, S. C.; Hariss, D. K.; Magnuson, V. R. POLLY; University of Minnesota: Duluth, MN, 1988.

(19) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1990**, *4*, 185.

(20) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581.

(21) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(22) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1990**, *4*, 185.

(23) Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press: 1986; pp 225−346.

(24) Hansch, C. *Acc. Chem. Res.* **1969**, *2*, 232.

(25) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.

(26) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.

(27) Dayhoff, J. *Neural Network Architectures*; van Nostrand Reinhold: New York, 1990; p 259.

(28) Khanna, T. *Foundations of Neural Networks*; Addison-Wesley: New York, 1991.

(29) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. *J. Comput.-Aided Mol. Design* **1994**, *8*, 405.

(30) Andrea, T.; Kalayeh, H. *J. Med. Chem.* **1991**, *34*, 2824.

(31) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. *J. Med. Chem.* **1994**, *37*, 3758.

(32) Peterson, K. L. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896.

(33) Zupan, J.; Gasteiger, J. *Anal. Chim. Acta* **1991**, *248*, 1.

(34) Lohninger, H. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 736.

(35) See e.g., (a) Ventura, S.; Silva, M.; Perez-Bendito, D.; Hervas, C. *J. Chem. Inf. Comput. Sci.* **1997**, *7*, 287−291. (b) Ni, Y. N.; Li, C.; Kokot, S. *Anal. Chim. Acta* **2000**, *419*, 185.

(36) Basak, S. C. *Med. Sci. Res.* **1987**, *15*, 605.

(37) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R.; Veith, G. D. *Math. Modelling* **1987**, *8*, 300.

(38) Hansch, C.; Leo, A. *Substituent constants for Correlation Analysis in Chemistry and Biology*; Wiley: New York, 1979.

(39) Hammett, L. P. *Physical Organic Chemistry*; McGraw-Hill: New York, 1970.

(40) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature* **1986**, *323*, 533.

(41) Basak, S. C.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 366.

(42) Trijanstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983; Vol. II.

(43) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Cheichester, U.K., 1983.

(44) Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31.

(45) Weiner, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.

(46) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *Toxicol. Lett.* **1995**, *79*, 239.

(47) Kshirsagar, A. M. *Multivariate Analysis*; Marcel Dekker: New York, 1972.

(48) Greenacre, M. J. *Theory and Application of Correspondence Analysis*; Academic Press: New York, 1984.

(49) Neural Works Professional-II/PLUS; NeuralWare, Inc.: Pittsburgh, PA, 1991.

(50) Documentation for Sybyl/HQSAR module is available from Tripos BookShelf website (http://www.tripos.com/services/bookshelf).

(51) Sybyl 6.6 version, available from Tripos Associates, St. Louis, MI.

(52) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1991**, *7*, 243.

(53) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. *J. Med. Chem.* **1995**, *38*, 1431.

(54) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. *J. Med. Chem.* **1996**, *39*, 3049.

(55) Viswanadhan, V. N.; Ghose, A. K.; Kiselyov, A.; Weinstein, J. N.; and Wendoloski, J. J. In *Combinatorial Library Design and Evaluation: Principles, Methods, Software Tools and Applications for Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, to appear.

(56) Cerius2 4.0; available from Molecular Simulations, Inc.: San Diego, CA, 1999.

(57) Shnur, D.; Venkatarangan, P. In *Combinatorial Library Design and Evaluation: Principles, Methods, Software Tools and Applications for Drug Discovery*; Ghose, A. K., Viswanadhan, V. N., Eds.; Marcel Dekker: New York, to appear.

CI000072+