Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds

Thomas M. Frimurer,*,† Robert Bywater,† Lars Nærum,† Leif Nørskov Lauritsen,† and Søren Brunak‡

MedChem Research, Novo Nordisk Park, Novo Nordisk A/S, DK-2760 Måløv, Denmark, and Center for Biological Sequence Analysis, Department of Biotechnology, Technical University of Denmark, DK-2800 Lyngby, Denmark

Received January 29, 2000

We have used a feed-forward neural network technique to classify chemical compounds into potentially "drug-like" and "non drug-like" candidates. The neural network was trained to distinguish between a set of "drug-like" and "non drug-like" chemical compounds taken from the *MACCS−II Drug Data Report* (MDDR) and the *Available Chemicals Directory* (ACD). The 2D atom types (of the full atomic representation) were assigned and applied as descriptors to encode numerically each compound. There are four main conclusions: First the method performs well, correctly assigning 88% of the compounds in both MDDR and ACD. Improved discrimination was achieved by a more critical selection of training sets. Second, the method gives much better prediction performance than the widely used "Rule of Five", which accepts as many as 74% of the ACD compounds but only 66% of those in MDDR, resulting in a correlation coefficient which is effectively zero, compared to a value of 0.63 for the neural network prediction. Third, based on a standard Tanimoto similarity search the selection of drug-like compounds in the evaluation set is not biased toward compounds similar to those in the training set. Fourth, the trained neural network was applied to evaluate the drug-likeness of 136 GABA uptake inhibitors with impressive results. The implications of applying a neural network to characterize chemical compounds are discussed.

INTRODUCTION

The design and discovery of drugs is a lengthy and costly process requiring a wide variety of technologies and a multitude of skills from computational analysis and synthetic chemistry to pharmacology, pharmacokinetics, metabolic studies, and clinical testing. The earliest phases of this process involve rational design or high throughput screening. In the latter case where automation is now routine, literally thousands of potential candidates can be screened. This high yield poses the question: can we improve the *quality* of the libraries we are making. Herein lies yet another problem because quality needs to be defined with reference to many different criteria such as (1) potency and efficacy, (2) ease of synthesis, (3) flexibility, (4) solubility, (5) uptake and distribution by the patient (dependent on the administration method), (6) drug metabolism and pharmacokinetics, (7) toxicity, and (8) stability, chemically as well as metabolic.

The continuous growth in drug discovery of combinatorial chemistry methods,[1,2] where large numbers of chemical compounds are synthesized and screened in parallel for in vitro pharmacological activity, has dramatically increased the need for rapid and efficient models for estimating the "drug-likeness" of the compounds. Many potential drug candidates which may look excellent in the laboratory can become expensive failures later when they go into clinical trials. Therefore the ability to effectively predict if a chemical compound is "drug-like" or "non drug-like" would be a valuable tool in the design, optimization, and selection of drug candidates for development.

In medicinal chemistry every effort is made to engineer high selectivity/specificity into drugs, but there may be some features that are common to all successful drugs. The reasons for this are likely to reside in the more general requirements mentioned above (absorption, uptake, resistance to rapid degradation etc.). A set of assumptions about necessary features for a "good" drug is embodied in the so-called "Rule of Five".[3] We emphasize that this rule was not designed for classifying molecules in the way we wish to do here. Nevertheless, it is often used, albeit erroneously, for this purpose and therefore we shall include an examination of how well it serves the purpose of discriminating drugs from nondrugs as defined in this paper and earlier papers cited herein[4−6] in our analysis.

There is a wealth of information implicitly encoded in the 2D- and 3D- dimensional structures of drugs currently sold and drug candidates in late development (i.e. chemical compounds in phase I, II, or III, preregistered, registered, preclinical, in clinical trials or launched), and it is our intention in this paper to examine ways of extracting and using these data.

Recently[5,6] neural networks were applied to address the question of whether a given chemical compound is "drug-like" or "non drug-like" with encouraging results. However, only a very small training set was used, less than 3% of the total, and, in the previous work a redundant data set was used, which may lead to a significant overestimation of the predictive performance.[7] Our neural network method embodies significant improvements. First we use a nonredundant

* Corresponding author phone: +45 44 43 45 10; e-mail: tfri@novo.dk.
† MedChem Research.
‡ Technical University of Denmark.

10.1021/ci0003810 CCC: $19.00
Published on Web 10/04/2000

data set. Redundancy *within* databases can augment the learning capabilities of the net, but redundancy *between* them in general introduces noise and reduces the discriminating power. Redundancy here means that the compounds in their vectors, component for component, are similar. Similarity to the neural network is a completely different issue, because the network can correlate all the vector components with each other in a nonlinear fashion. Therefore two very different compounds may appear very similar to the neural net. However, it is clear that redundant data will constrain the network weight structure less than nonredundant data. Nonredundant data will therefore in most cases lead to well-performing networks even if the weights to training examples ratio is larger. Second, we used in the order of 75% of the available data for training the neural network. This was done to simulate a real selection procedure of industrial relevance where one normally is interested in filtering out a few true positives (i.e. "drug-like" compounds) among hundreds of thousands of compounds in a virtual library. Third, the large number of neurons in the hidden layer of the neural network used here is a requirement for handling a large amount of highly nonlinear input data. Earlier authors used very small networks with less than 10 hidden units.[5,6] Fourth, we address the issue of designing the most important descriptors.

METHODS

Databases. In this study the *MACCS−II Drug Data Report* (MDDR)[8] and the *Available Chemicals Directory* (ACD)[9] act as a surrogate for "*drug-like*" and "*non drug-like*" molecules. MDDR and ACD contain ∼99,000 and 250,000 compounds, respectively. The MDDR and ACD molecular structures were filtered according to the following rejection criteria:

(i) All compounds lacking the connectivity table or containing obvious errors in the structure description or errors detected by CONCORD[10] in the two databases were removed within ISIS.[11]

(ii) All counterions (Cl^-, Na^+, K^+ etc.) and solvent molecules were removed from the structure field.

(iii) Identical compounds found in both databases were removed from ACD.

(iv) In principle it is possible to extract any desired subset of MDDR as a training set, but for our purposes we removed compounds such as minerals, vitamins, aerosol propellants, spermicides, cytostatika, radio-photo sensitizers, ultraviolet light absorbers, blood components/substitutes, vaccines, antibacterial, -viral, and -protozoal agents, antibiotics of various classes: penicillin, macrolide, aminoglycoside, tetracycline, etc., metal complexes: platinum, tin, gold, ruthenium, rhodium etc. The identity number of the compound groups in these obmitted categories are (38100, 38200, 49200−02, 59120−30, 59820, 60100−61000, 64100−71000, 71580, 75200−75400, 75835−77000, 79000−80501, 80510−55, 81224−83000, and 99000). Out of a total of 99,000 compounds, 26,000 were removed.

The MDDR compounds can be divided into two categories: (1) a small and diverse collection of compounds which all are in clinical trials i.e., Phase I, Phase II, Phase III, preregistered, registered, and finally launched, and (2) a much larger group comprising all the remaining compounds having the status "Biological Testing". In this study the first group of MDDR compounds acts as "real drugs", while the second acts as "drug candidates". After applying the above rejection filter, these groups contain ∼4500 and ∼68,500 compounds, respectively. The first group of MDDR compounds was applied for training and testing, the second group only for testing. Applying the filters (i), (ii), and (iii) to ACD leaves 169,000 compounds.

To obtain as nonredundant a data set as possible the 4500 MDDR compounds (classified as real drugs) were subjected to a diversity filter applying the ISIS fingerprints[11] and a Tanimoto limit[10] of 0.85, i.e. all compounds having a Tanimoto coefficient exceeding 0.85 to the 4500 MDDR compounds were removed. This reduces the number of compounds by around 100 leaving in the order of 4400 compounds for training and testing. Applying the same similarity filter to all MDDR compounds bearing the labeling "Biological Testing" reduces this group by 8500 compounds leaving 60,000 for test purposes. It is assumed that ACD contains compounds which mainly are "non drug-like", but a small fraction of this database may in fact be drugs. To avoid false classification as much as possible all close analogues of the diverse 4400 MDDR molecules were removed from ACD by applying to these compounds a second standard Tanimoto similarity search using the ISIS fingerprints and a cutoff level of 0.85. By applying this similarity filter 79,000 ACD compounds were removed, leaving 90,000 diverse ACD compounds for training and testing. We are well aware that certain "obvious" nondrugs such as reagents had not been removed from ACD. Even if certain nondrugs (e.g. reagents) may be identified by simple rules in the form of search for specific functional groups, the network may learn correlations from them which can have a positive effect on the classification of other compounds.

Two Training and Test Sets. Data Set 1. The neural network was trained on randomly selected compounds from ∼75% of MDDR (all classified as "real drugs") and from ACD. The remaining ∼25% of the two databases were used for testing. The training set consisted of 3000 MDDR and 70,000 ACD compounds, while the test set contained 1,400 MDDR and 20,000 ACD compounds, respectively. The training was performed with and without a balancing procedure.[16] The ratio between the number of MDDR and ACD compounds is about 1:10, i.e., there are 10 times as many ACD (non drug-like) compounds available compared to MDDR (drug-like) compounds. Therefore, in order not to make the network over-predict nondrugs, the training included a balancing procedure with additional (randomly selected) presentations of the training examples in the drug category. The 60,000 MDDR drug candidates having the label "Biological testing" were evaluated separately.

Data Set 2. For comparison**,** we created this data set along the same lines as earlier reported,[5,6] whereby small training sets (5000 and 3500, respectively) were obtained by random partitioning of the ACD and MDDR. Our version of this data set contained 4400 randomly selected compounds from ACD and from MDDR of which 3000 from each were used for training and the remainder for testing.

All the MDDR and ACD compounds were assigned a score value of 1 and 0, respectively. When interpreting the network output, we used a threshold of 0.5 to discriminate

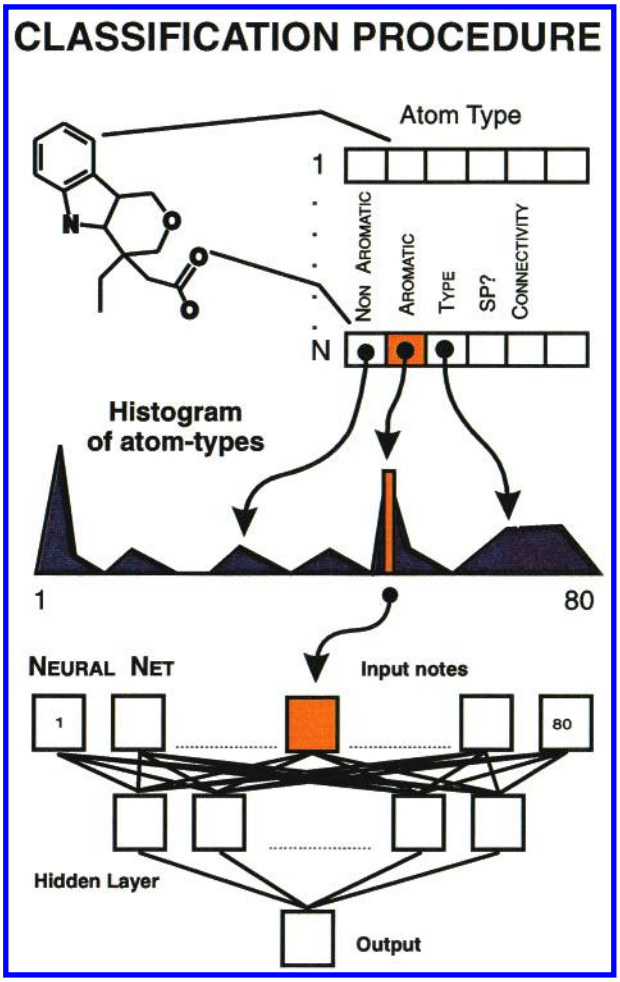

Figure 1. Illustration of the classification procedure. The chemical compounds are encoded numerical using atom types as descriptors which are used as input for training and testing the neural network.

positives i.e., "drug-like" and negatives i.e., "non drug-like" compounds.

Molecular Descriptors. The classification procedure is illustrated in Figure 1. For each compound, a string of real numbers normalized between 0 and 1 representing the atom type distribution is generated. The normalization was carried out with respect to the maximum of each descriptor. (The array or histogram size required to store the descriptor vector representing the compounds is invariant with respect to compound size and the order in which the atoms are read). The atom-types were used as descriptors for encoding numerically the MDDR and ACD chemical structures.

The program CONCORD[10] was used to assign atom types and molecular descriptors for the compounds. Each atom may be assigned as many as six atom-type codes, each providing additional detail concerning the chemical context in which the atoms appear. A similar atom-type coding has been successfully employed in the PRODRG molecular descriptor method.[12] A total of 80 molecular descriptors were generated for each compound. The 2D atom type descriptors encoded by the CONCORD software package are given in Table 1. All the chosen descriptors are populated in the training and test set of both the ACD and MDDR compounds. The mapping process of atom-types (from an array of the CONCORD atom type codes) was done within our own user defined subroutine compiled and linked to the CONCORD object file library.

Table 1. CONCORD Atom Types Descriptors Applied as Input to the Neural Net[a]

HYDROGEN:	CARBON:	NITROGEN:	OXYGEN:	PHOSPHOUR:	FLUORINE:
X-H aro.*	Aromatic.	Aromatic	Non-arom. *	Non-arom.	any*
X-H non aro.*	Non-arom.	=N- *	sp2 (any) *	sp3 (any)	
C-H (any) *	sp	non-arom. *	O=C		CHLORINE:
N-H any	#C-	sp (any)	O=N	SULPHUR:	any *
>N-H	=C=	sp+	O=P		
O-H	sp2	sp2 (any)	O=S	Non-arom.	BROMINE:
O-H	C=C	N=C *	-O-	sp2*	any
=CO-H	C=N	N=N	O -1 charge	=S	
C=CO-H	C=O *	N=O	O=X=	-S-	IODINE:
C(=O)O-H	C=S	-N<	C-O-C	sp3	any
S-H	amide C	amide N	C-O-H	-S-	
S-H	C=X=	vinyl N	sp3 (any)	three subst.	OTHER:
	sp3	-N=X=		O=S<	
	sp3-H	=[N+] <	SILICON:	four subst.	No. atoms *
	sp3-H_2	=[N+] < conj.	Non-arom.	sulfones	heavy atoms
	sp3-H_3	sp3 (any)	sp3 (any)		total charge *
	sp3 in cyclop.	> [N+] <			
	sp2 conj *				

[a] The fifteen most important descriptors are indicated by asterisks (see text).

The MDDR and ACD compounds containing the atom-type descriptors listed below appear so rarely in the nonredundant data set that the neural network cannot exploit them for generalization. Therefore compounds which contained one or more of the these atom-type descriptor were removed from the subset:

```
#CO-H, Si-H, P-H
C=Si, C=P, Sp2 in cyclopropanyl, Sp2+
=[N+] <, >N-, N=Si, N=P, N=S
Aromatic O, O=O, O=Si
Silicon Sp or Sp2
Phosphorus Sp Sp2 or P with three substituents
Aromatic Sulfur Sp or Sulfur with three substituents
```

Neural Network Algorithms. The Neural Network as a Flexible Modeling Method. Multilayered feed-forward neural networks with error back-propagation have been successfully applied to several practical classification problems in chemical spectroscopy.[13,14] In computational and medicinal chemistry neural networks have been employed in quantitative structure−activity relationships[15] and in scoring for molecular similarity.[16,17] Within the field of bioinformatics there are numerous successful applications; for reviews see refs 7 and 18.

In general, neural networks are capable of representing the nonlinear relation between variables due to the complex connections among the neurons and the flexibility in using processing functions. The neural network applied in this study was the HOWLIN program developed at the Centre for Biological Sequence Analysis, Department of Biotechnology, Technical University of Denmark. The architecture has been reviewed elsewhere in a number of articles.[19−22]

The basic neural network model considered here has an input layer, one hidden layer, and a single output unit with logistic activation. Each unit besides those in the input layer calculates a weighted sum of its input and passes this sum through a nonlinear sigmoidal function to produce the output. The output response of a unit when presented by a vector input I is given by

$$O = \sigma\left(\sum_{n=1}^{N} w_n \mathrm{I}_n - t\right) \qquad (1)$$

The weights w_n are the adjustable parameters of the network and are through a training process modified iteratively to produce desired output for given input vectors. σ is a sigmoidal function of the type $\sigma(x) = 1/[1 + exp(-x)]$, and

t is a threshold parameter. The objective of the training procedure is to find a set of weights such that the network will produce output values as similar as possible to the known targets. This is done by minimizing an error function according to the method described by Rumelhart[23]

$$E = \sum_{\alpha;i} (O_i^\alpha - T_i^\alpha)^2 \qquad (2)$$

where α enumerates the examples.

A practical issue that arises in the use of neural networks concerns their generalization capability. Is the trained neural network able to perform good predictions in samples different from the training set? The answer to this question is not easy to formulate, and methods to improve generalization have recently been developed based on different assumptions, such as network pruning approaches.[24,25] By minimizing the complexity of the network better results can be obtained. A neural network with minimum size is less likely to learn the idiosyncrasies or noise in the training data and may thus better predict on new data.

Performance Evaluation of the Model. The performance of the model studied is presented in terms of the standard Mathews correlation[26] defined as

$$C = \frac{(PN) - (N^f P^f)}{\sqrt{(N + N^f)(N + P^f)(P + N^f)(P + P^f)}} \qquad (3)$$

Here P and N are the number of true positive and true negative predictions. P^f is the number of false positives, and N^f is the number of false negatives. For a perfect prediction the correlation coefficient is equal to 1, and for a completely erroneous prediction it is equal to −1.

Conceptually, the Pearson correlation[27] is the ratio of the variation shared by N and P to the variation of N and P separately. The formula is

$$r = \frac{\sum (N - \bar{N})(P - \bar{P})}{\sqrt{\sum (N - \bar{N})^2} \sqrt{\sum (P - \bar{P})^2}} \qquad (4)$$

where the nominator is the sum of products of corresponding deviation scores for two variables while the denominator is the square root of the product of the deviation scores for the two variables. When there is a perfect linear relationship, every change in the N variable is accompanied by a corresponding change in the P variable. In this case, all variation in N is shared with P, so the ratio given above is $r = 1.00$. As usual $\bar{N}$ and $\bar{P}$ represent the mean. At the other extreme, when there is no linear relationship between N and P, then the numerator is zero, and $r = 0.00$.

Neural Network Training. To avoid over-fitting and to guarantee generalization ability, a validation set is used to compute the cost function. The two most important methods to avoid over-fitting (often called "over-training") are regularization methods (such as weight decay) and early stopping. The regularization methods try to limit the complexity of the network such that it is unable to learn single example peculiarities. Early stopping aims at terminating the training at the point of optimal generalization. We applied the latter strategy by stopping the training at maximal Matthews correlation obtained for the test set. The correlation learning architecture is an example of this approach.

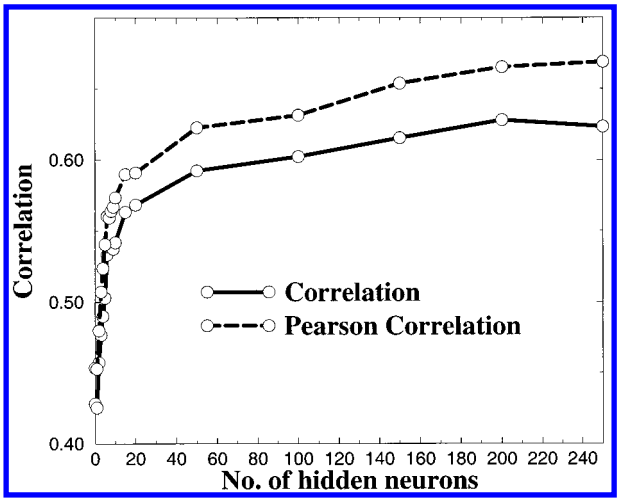

Figure 2. Predictive performance of the neural network with variable number of neuron in the hidden layer. The performance is quantified by the correlation coefficient C calculated from (eq 3). It is seen that the optimal performance is obtained for a neural network having 200 neurons in the hidden layer. The performance of the neural network increases with an increasing number of hidden units due to the nonlinearity of the chemical compounds which the networks implicitly must learn.

RESULTS AND DISCUSSION

Optimal Neural Network Configuration for Data Set 1. The optimal network configuration was found by training and evaluating a number of different neural network architectures. A total of 18 neural network configurations having 0−10, 15, 20, 50, 100, 150, 200, and 250 neurons in the hidden layer were trained and evaluated. The predictability of each configuration (on the test sample i.e., 1400 MDDR compounds all in clinical trial i.e., Phase I, Phase II, Phase III, preregistered, registered, and finally launched compounds) and 20,000 compounds from the ACD database was evaluated by calculating the correlation coefficient (eq 3). The predictive correlation on the test set as function of neurons in the hidden layer is shown in Figure 2. The optimal performance is reached by having 200 neurons in the hidden layer giving a learning and test Mathews correlation coefficient of C = 0.65 and C = 0.63, respectively. The network reaches maximal test performance after 605 learning cycles. This is also the point where the neural network has its optimal generalization capability. It is clear that the relationship between the physicochemical properties and the final biological activity is highly nonlinear, since the neural network performance increases with increasing number of hidden units.

Figure 3 shows a normalized distribution of all the ACD (solid line) and the MDDR (dashed line) scores. In addition, the predicted score distribution of the 60,000 MDDR compounds bearing the label "Biological Testing" is shown (long dashed line). The neural network trained on data set 1 predicts 50% of the MDDR compounds bearing the label "Biological Testing" to be drug-like molecules applying a threshold of 0.5. The significantly lower predicted content of "drug-like" chemical compounds is expected since the MDDR molecules in this case are to be considered as drug candidates rather than real drugs.

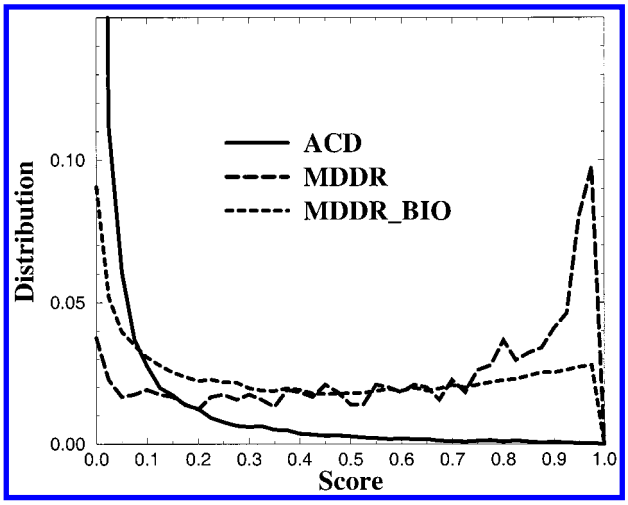

Figure 3. Distribution of predicted score value for 1400 MDDR and 20,000 ACD test compounds. The *solid* and *long dashed line* represents the distribution of the ACD and the MDDR compounds, respectively. For threshold 0.15 it correctly classifies 88% of both. Also the distribution of the predicted score value for 60,000 MDDR (all bearing the label "Biological testing") *dashed line*. The network trained on data set 1 predicts that 50% of the MDDR (all bearing the label "Biological testing") are drug-like.

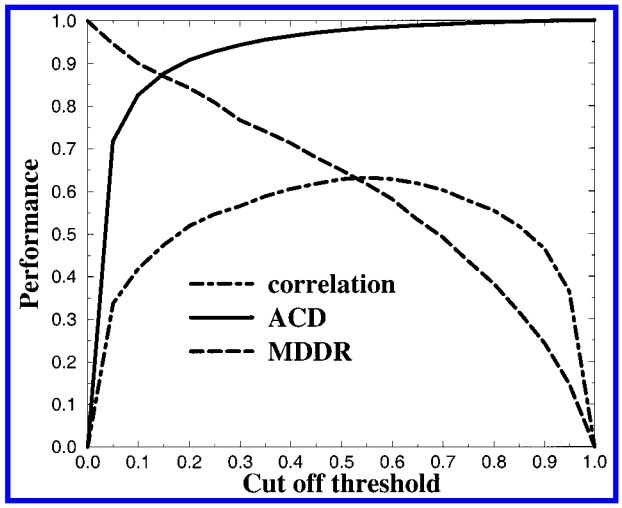

Figure 4. The figure shows the performance correlations (*dashed−dotted line*), the ratio of correctly predicted to total ACD (*solid line*), and likewise for MDDR (*long dashed line*) compounds as function of acceptance value for "drug-like" versus "non drug-like") for data set 1 (see text). For an acceptance value equal to 0.5 more than 98% of the ACD 63% and of the MDDR compounds are classified correctly. Lowering the acceptance value increases the number of correctly predicted MDDR compounds but introduces an increased number of false positive predicted ACD compounds. For an acceptance value of 0.15, 88% of the MDDR and ACD compounds are classified correctly.

As mentioned earlier we wish to obtain a real filtering procedure that selects a few "true positive" compounds among hundreds of thousands of chemical compounds. In Figure 4 the performance correlation (dashed − dotted line) and the ratio of correctly predicted ACD (solid line) and MDDR (long dashed line) compounds as a function of the threshold value for "drug-like" versus "non drug-like are shown. For a threshold value of 0.5, more than 98% of the ACD and 63% of the MDDR compounds are classified correctly. Lowering the threshold value increases the number of correctly predicted MDDR compounds but introduces an increased number of false positively predicted ACD compounds. For a threshold value of 0.15, 88% of the MDDR and ACD compounds and ∼75% of the ("Biological Testing") subset from MDDR are classified correctly.

Optimal Neural Network Configuration for Data Set 2. As before, 18 neural network configurations having 0−10, 15, 20, 50, 100, 150, 200, and 250 neurons in the hidden layer were trained and evaluated. This time a network with 9 neurons in the hidden layer was found to be optimal. This can be compared with the 5 and 5−10 hidden units applied in the two recent studies.[5,6] Learning capacity for training was 92.5% giving a learning correlation of 0.86. Total learning performance was 87.5% (test correlation 0.75), with 86% correctly predicted in the ACD and 89.6% in the MDDR. Figure 5 shows the neural network performance correlation for data set 2. The line types are the same as in Figure 4. At a cutoff threshold of 0.5, 86% of ACD and 87% of MDDR compounds were correctly predicted giving a performance correlation C = 0.75. At a threshold of 0.85, 97% ACD and 56% MDDR were correctly assigned.

Comparison of Results for Data Sets 1 and 2. The differences between the performance with the two data sets can be seen by comparing Figure 4 with Figure 5**.** The discriminating power is clearly superior with the larger data set 1. For example if one is interested in effectively avoiding false negative predictions (i.e non drug like compounds) a threshold of 0.35 and 0.75 can be used for the neural networks trained on data sets I and II, respectively. Using these thresholds the neural networks correctly assign 95% of the ACD compounds in both test sets but eliminate about one-fourth of the MDDR when trained on data set 1, while almost 30% of the MDDR compounds are discarded when trained on data set 2.

The performance curves can be summarized in the following table:

data set	cutoff	ACD correctly predicted, %	MDDR correctly predicted, %
I	0.05	72	95
II	0.3	69	95
I	0.35	95	74
II	0.75	95	70

Results of Reduced Descriptor Set. It is known that by minimizing the complexity of the network sometimes better results can be obtained. A neural network with minimum size is less likely to learn the idiosyncrasies or be influenced by noise in the training data and may thus better predict on new data. The importance of the individual descriptors was estimated by applying the neural network trained to score all the test compounds, each time leaving one descriptor out by setting the respective descriptor value equal to 0. The impact of each of the 80 descriptors is thus evaluated when the remaining 79 are still present in the input. Larger errors may in some cases arise if two, three, or four (and so on) descriptors would be removed simultaneously; this would however require testing of hundreds or thousands of networks. In Table 2 a list sorted according to the size of the error (eq 2) is presented. We systematically reduced the descriptor set by removing less important descriptors. In this way 65 out of 80 descriptors could be removed from the data set. In general, it is hard to reduce the number of descriptors in a unique database without increasing bias i.e., the number of duplicates within both databases and between the databases. Using these 15 important descriptors to

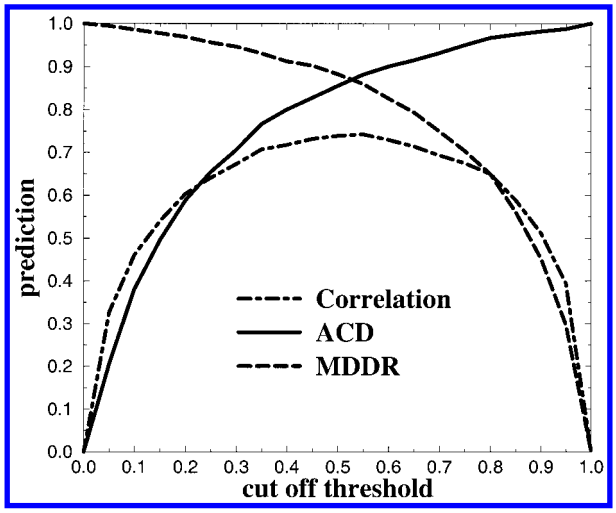

Figure 5. The figure shows the performance correlation for data set 2 (see text). The line types are the same as in Figure 3. At a cutoff threshold 0.5, 86% of ACD were correctly predicted and 87% of MDDR with correlation C = 0.75. At threshold 0.85, 97% ACD and 56% MDDR.

characterize the compounds resulted in 2128 "identical" ACD compounds and 9 identical MDDR compounds. Surprisingly, only 21 compounds were identical between the MDDR and ACD databases. These 21 compounds were removed from the ACD training and test sets. Applying the 10 most important descriptors resulted in a significant overlap between the MDDR and ACD compounds.

Applying data set 1, a neural network was trained on the 15 most important descriptors (**X−H arom, X−H non arom, C−H (any), C═O, C Sp2 conj, ═N-, N nonarom, N═C, O Nonarom, O Sp2 (any), P Sp2, F any, Cl any, number of atoms and total charge,** also highlighted in Table 1 by asterisks) found by the procedure described above. As previously, 16 neural network configurations having 0−10, 15, 20, 50, 100, 150, 200, and 250 neurons in the hidden layer were trained and evaluated. The network reaches its maximal test correlation C = 0.54 after 2701 learning cycles having 100 hidden units. Applying a cutoff threshold of 0.5, the trained neural network correctly predict 97% of the ACD and 45% of the MDDR compounds. Using a cutoff threshold of 0.15, 87% of the ACD and 82% of the MDDR compounds were correctly predicted. Thus by reducing the 80 descriptors to 15 it is still possible to train a neural net to only slightly reduced performance. The relatively high performance of the 15 descriptor network shows that we have identified 15 of the most information rich descriptors. Finaly, different cross validation calculations have been attempted without any significant change in correlation performance.

Evaluation of the Model. The reliability of the neural network performance was estimated by the following: (1) comparison of the neural network prediction to the Rule of Five, (2) behavior of the prediction, that is how similar (in terms of Tanimoto similarity) are the predicted compounds in the test set to compounds in the training set, (3) a negative test in which 10,000 randomly chosen ACD compounds were split into two groups, one assigned as non drug-like and the other as drug-like to see if there is a real difference between the two database parts, (4) individual expert assessment of suspected drug-like ACD compounds, (5) 136 GABA-uptake inhibitors synthesized and biological tested in house were examined for drug-like character, (6) NCI compound scores, and (7) and last test on recent additions to MDDR. Since the above results were obtained a further 226 MDDR compounds have been added to the MDDR.

Table 2. Importance of the Individual Atomic Descriptors Was Estimated by Analyzing the Performance of the Network Trained on Data Set 1 (by Setting the Respective Descriptor Equal to 0) and Then Observing the Error Introduced by the Descriptor Mutation[a]

overall error	atom type descriptor	overall error	atom type descriptor
1043.44	no. of heavy atoms	1613.84	C═CO−H
1138.09	sp2 (C any)	1614.54	O═N
1153.73	aromatic C	1615.47	O═P
1184.23	nonarom C	1615.48	sp3 in cyclop. (C)
1209.76	sp2 (N any)	1616.49	amide N
1283.21	O═C	1620.43	S-H
1304.32	Sp3 (C any)	1623.21	three substituents S
1354.00	sp3-H3 (C)	1624.60	Sp (N any)
1386.52	sp3-H2 (C)	1626.00	iodine any
1446.90	C═N	1631.32	Sp (C any)
1474.33	Sp3 (N any)	1633.49	>N−H (sp3 any)
1513.07	sp3-H (C)	1638.12	−S-
1540.72	N-H any	1642.26	vinyl N
1554.42	O-H	1645.27	sp3 (sulfur any)
1557.93	sulfones	1672.19	N═N
1564.36	−O−H	1674.00	═[N+]<
1574.83	C-O−H	1674.15	C-O−C
1584.8	═CO−H	1684.34	bromine any
1587.04	−S-	1688.29	C═C
1587.27	C(═O)O−H	1695.33	Sp3 (O any)
1588.68	C═S	1703.39	═S
1596.94	four subst. S	1710.75	nonaromatic sulfur
1604.03	═[N+] < conj.	1730.59	O═S
1604.20	nonaromatic P	1813.90	−N<
1605.22	Sp3 (P any)	1843.20	aromatic N (any)
1606.46	> [N+] <	1864.40	Sp2 (sulfur any)
1607.25	−S−H any	1880.81	═N-
1607.75	N═O	2001.92	chlorine any
1608.81	−N═X═	2018.75	Sp2 (O any)
1608.90	C═X═	2112.10	nonaromatic O
1608.94	═C═	2121.63	fluorine any
1609.68	nonaromatic silicon	2426.52	X-H aromatic H
1609.68	Sp3 (silicon any)	2688.46	N═C
1609.91	sp+	2940.40	charge
1610.82	O═X═	3210.59	Sp2 conjugated (C)
1610.91	O═S<	3217.12	nonaromatic N (any)
1611.31	#C-	3505.32	C═O
1611.66	amide C	3506.64	no. of atoms
1612.26	−O-	3711.48	X-H nonaromatic H
1612.67	O −1 charge	7452.42	C-H

[a] This was done for all 80 descriptors one by one. The largest ranked error was caused by zeroing the 15 descriptors indicated in blue.

(1) Comparison to the "Rule of Five". The "Rule of Five" procedure has attracted widespread attention as a method of discarding those members of a compound library that are outside the limits of what is regarded as suitable from the point of view of bioavailability, considered in terms of complexity, hydrogen bonding potential, and excessive lipophilicity. These are features which are all important in evaluating drug-likeness. Although the Rule of Five is not a "drug/non drug" discrimination method as such, we report here the results of testing the rule in this role.

The *minimum* and *maximum* default parameters involved in the compound selection process are the following:

```
i)       0 <= number of donor groups     <= 5
ii)      0 <= number of acceptor groups  <= 10
iii) -9.99 <= MlogP                      <= 4.15
iv)      0 <= molecular weight           <= 500.
```

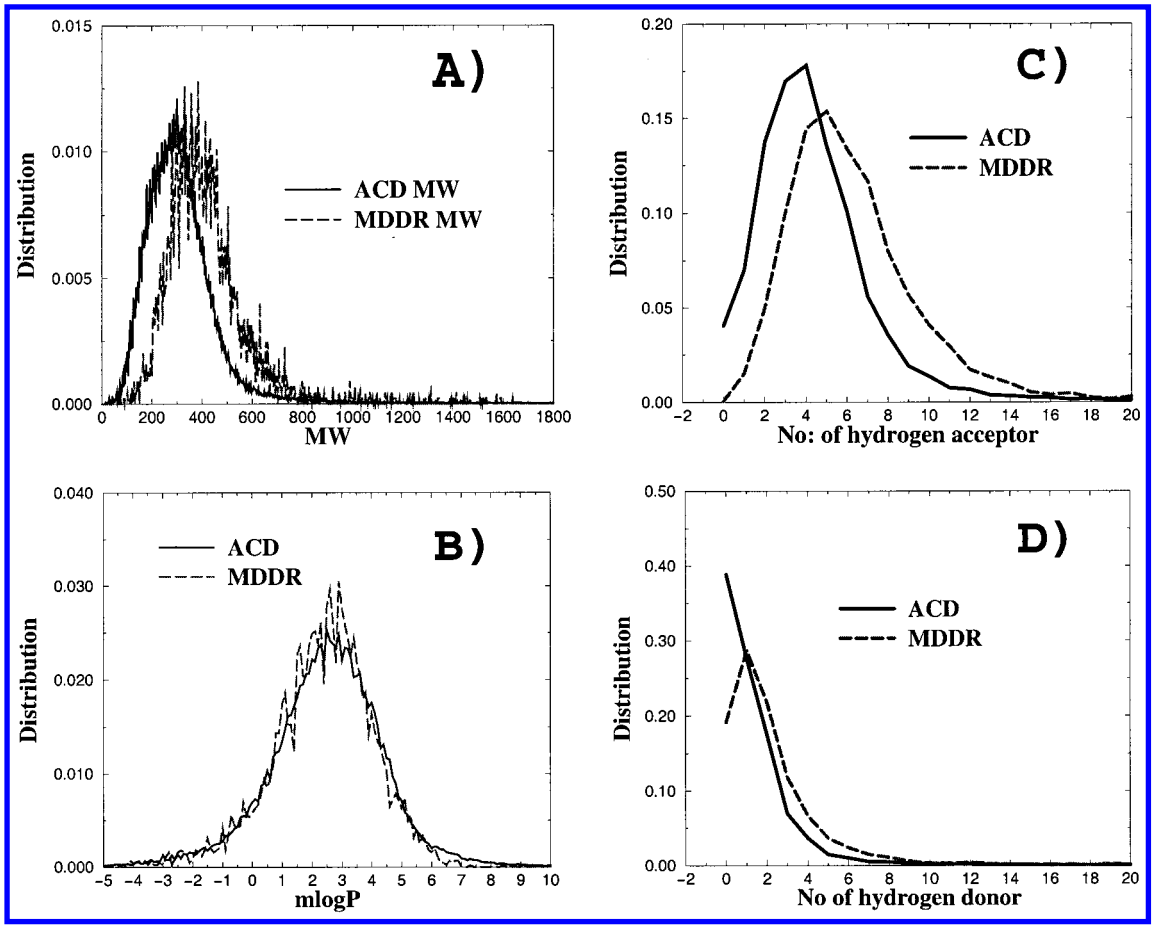

Figure 6. A) and C) indicate a shift toward higher average MWs and numbers of acceptor groups for the MDDR relative to the ACD compounds. The distribution of calculated MlogP values B) and number of donor groups D) are identical in the ACD and MDDR compound databases.

A ClogP value of 5.00 (equivalent to an MlogP value of 4.15) constitutes a general cutoff for oral availability.[3] To fulfill the Rule of Five a compound must fit all four ranges (i.e. logical AND) to be accepted. It is expected that any method that distinguishes "drug-like" molecules from "non drug-like" molecules should classify the major part of the MDDR compounds as drug-like. The result of applying the Rule of Five to all the MDDR and ACD compounds shows that 66% of the MDDR compounds were accepted, while 34% were discarded. In contrast to this 76% of the ACD compounds were accepted, while only 24% of the ACD compounds were discarded i.e., classified as non drug-like. Applying the correlation expression given by eq 3 gives the correlation $C = -0.034$, very close to random assignment.

The distribution of the "Rule of Five" selection criteria for all the ACD and MDDR compounds is shown in Figure 6. A) and C) indicate a shift toward higher averages MW and numbers of acceptor groups for the MDDR relative to the ACD compounds. A shift toward higher MW for compounds in clinical trial has also been shown by Lipinski.[3] The distribution of calculated MlogP values B) and number of donor groups D) are identical in the ACD and MDDR compound databases. It is apparent that the selection criteria for the Rule of Five significantly overlap the two databases.

It has been suggested by one of our referees that the Rule of Five should be modified to a MW range of 200 to 600 and that it should give an alert only when two or more of the four rules are violated. We found that when only compounds in the 200−600 MW range were tested, the Rule of Five accepted 79.9% MDDR and 77.9% ACD compounds. With two violations 98.0% were accepted from MDDR and 98.7% from ACD. The correlation for the two cases are 0.025 and −0.027, respectively. Hence even under conditions that should favor this rule, it is not ideal for obtaining a good discrimination.

(2) Comparison to the Standard Tanimoto Similarity. All ACD compounds which were classified as "drug-like" in data set 1 have been compared separately to all MDDR compounds in the training set in terms of a standard Tanimoto similarity. Figure 7 shows the distribution of ACD compounds "predicted to be drug-like by our model" as function of Tanimoto coefficients. The distribution peaks at a Tanimoto index value of 0.45, showing that the majority of the ACD compounds classified as drug-like by the neural net in terms of the Tanimoto metric is diverse compared to the MDDR compounds.

(3) Negative Test. To determine if there are real differences between the ACD and MDDR chemical structures we constructed training and test sets from ACD consisting of 10,000 randomly selected compounds. Half of the data was labeled as "non drug-like" with a score equal to 0 and the other half "drug-like" i.e., score equal to 1. From the data set 6000 compounds were applied for training and 4000 for test. During the neural network training a learning correlation of 0.1 could be achieved (corresponding to a 57% and 43% separation), while the correlation on the test set was 0. This

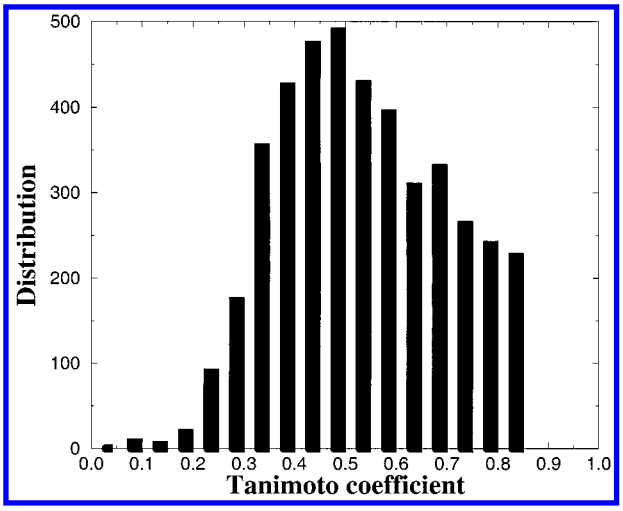

Figure 7. Histogram of Tanimoto coefficients. All ACD compounds which were classified as "drug-like" in the test set have been compared severally to the MDDR compounds in terms of a standard Tanimoto similarity. The figure shows the distribution of ACD compounds "predicted to be drug-like by our model" as a function of Tanimoto coefficients. The distribution peaks at 0.45 showing that the major part of the ACD compounds classified as drug-like by the neural net in terms of a Tanimoto metrics are diverse compared to the MDDR compounds.

showed that the original training on "correctly" classified data could pick up some real differences between the ACD and MDDR databases and therefore capture differences between "drug-like" and "non drug-like" molecules.

(4) Expert Assessment of Classified Drug-like ACD Compounds. All ACD compounds in the training set (20,000) predicted by the neural net to be drug-like were subjected to individual (human) expert assessment. In total there were 402 compounds having a score >0.5 and 152 of those were identifiable as being drug-like by expert assessments of two medicinal chemists. The most easily recognized drug-like analogues of known drugs were for example 1,3-diethyl-2-thiobarbituric acid, 9,10-dihydrolysergol, and cimetidine, which are not contained in the MDDR database.

(5) A Database of GABA Compounds. An example of how the neural net might function in practice was the case of a set of 136 GABA-uptake inhibitors which were examined for drug-like character. These compounds which have been synthesized and biologically tested in house are obtained from our internal confidential library, and none of the compounds are contained in the MDDR database. It was shown that they were predominantly drug-like (Figure 8A), furthermore there was a trend in the relationship between activity and predicted drug-like character (Figure 8B). The majority of these compounds is predicted to share high drug-like scores > 0.5. These compounds have indeed good biological activities. On the other hand, many of the compounds predicted to be non drug-like (i.e. having scores < 0.5) have rather bad biological activities.

(6) NCI Compound Scores. The definition of what is regarded as drug-like in terms of marketed therapeutic preparations raises a number of questions. Certain categories of compounds such as antibiotics, antiparasitics, and cystostatika are designed to have properties, in regards to toxicity for example, that are likely to distinguish them from

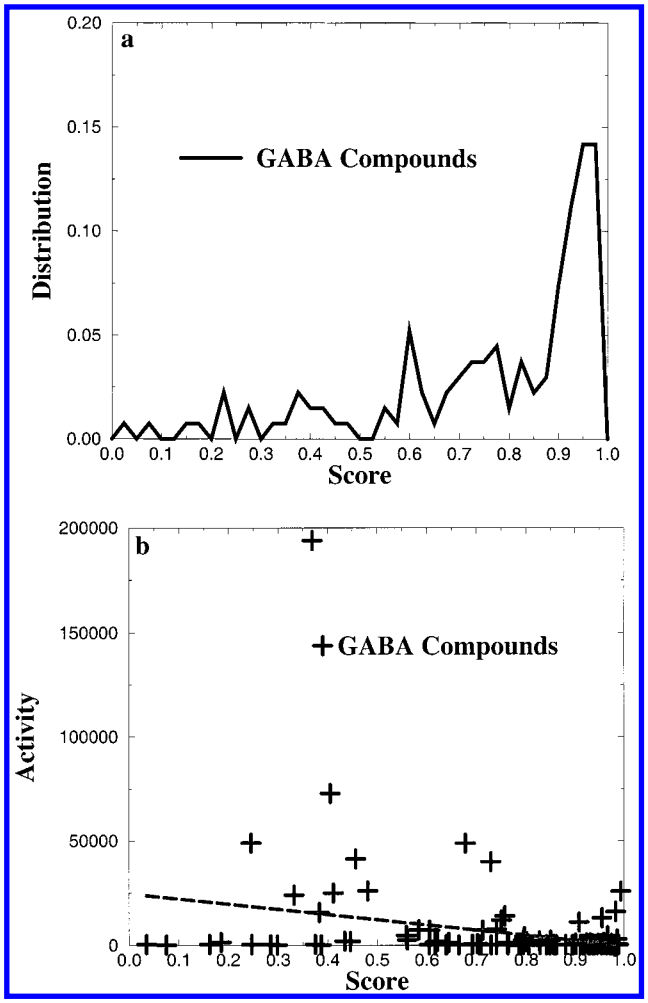

Figure 8. A) shows the neural network score distribution for 136 GABA-uptake inhibitors of which 84% are classified as drug-like. B) shows the clustering of compounds with high activity and predicted drug-like character.

compounds designed to mimic the action of endogenous substances in the patient such as neurotransmitters, enzyme substrates, or cofactors etc. Without any detailed analysis as to whether the compounds were particularly carcinogenic or toxic in the sense of having cytostatic properties, we examined the databases of compounds distributed by the National Cancer Institute NCI. As shown in Figure 9 the majority of these compounds has very little drug character, less than 1.5% could be classified as being drug-like, once again validating the approach present in this work.

(7) Test on Recent Additions to MDDR. Since the above results were obtained a further 226 compounds have been added to the MDDR. These are completely new compounds that have never been used for training or testing. Furthermore these compounds (except for few) are not similar to the MDDR compounds in our training set. It was shown that 85% of these were classified as drug-like with a threshold of 0.15.

CONCLUSION

We have developed an artificial intelligence tool which effectively discriminates between "drug-like" and "non drug-like" chemical compounds. Our method correctly assigns 88% of the MDDR and ACD compounds by applying a cutoff threshold of 0.15. To effectively avoid false negatives

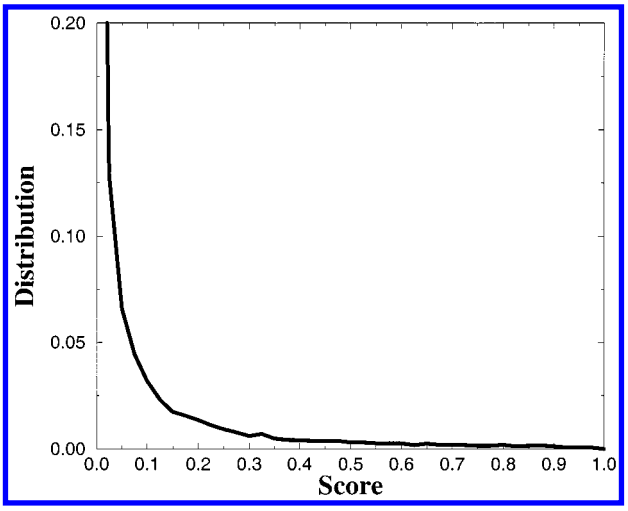

Figure 9. Neural network score distribution of 32,000 NCI compounds showing that the majority of these compounds have very little drug character, less than 1.5% could be classified as being drug-like.

of predicted drug-like chemical compounds a threshold of 0.5 can be used. In this case our model correctly predicts more than 98% of the ACD compounds and 63% of the MDDR compounds. We show that novel drug-like compounds can indeed be selected, and these are distinct from those selected on the basis of Tanimoto similarity using the ISIS fingerprints. Expert human assessment of ACD compounds predicted by the neural net to be drug-like and the drug-likeness of some hundreds of GABA uptake inhibitors synthesized and biological tested in house yielded comparable results.

The neural network method is very fast. The rate-limiting step is to encode the chemical compounds by the 2D atomic descriptors. From an input in SD format, the drug-likeness of 100,000 compounds/hour can be assigned on an SGI R10000 computer.

We do not expect that a simple rule could be used for discriminating between "drug-like" versus "non drug-like" chemical compounds. The large number of neurons in the hidden layer is a requirement raised by the need to handle a large amount of highly nonlinear input data. In contrast to earlier authors who used 5−10 neurons in the hidden layer, we had 200. In our case we applied 75% of the data set for training resulting in a modest improvement in performance and an increase in discriminating power. Furthermore we managed to reduce the descriptor set from 80 to 15 without increasing the bias and with little effect on accuracy. Finally, we show that our neural network method selects new drug-like compounds with a broad diversity compared to "drug-like" compounds in the original training set.

The result of applying the Rule of Five on all the MDDR and ACD compounds showed that 66% of the MDDR compounds were accepted, while 34% were discarded. In contrast to this 76% of the ACD compounds were accepted, while only 24% of the ACD compounds was discarded resulting in a correlation of $C = -0.034$. Thus, there is no correlation between the results of Rule of Five and "drug-likeness" as previously defined[4−6] and those reported herein.

The neural networks can be used together with combinatorial methods to generate artificial chemical compounds and to propose completely new "drug-like" chemical compounds. This should have a decisive effect for improving the quality of compound libraries.

Furthermore we are working on collecting databases containing chemical compounds with a specific pharmacological or pharmacokinetic profile in order to train the net for specific purposes such as (a) peptide/peptoid compounds as opposed to compounds which do not possess peptide-mimicking properties, (b) compounds with long or short (as desired) metabolic half-lives, and (c) scoring for toxicity or carcinogenicity.

ACKNOWLEDGMENT

We thank Knud Erik Andersen and Søren Padkjær for helpful discussions and computer assistance.

REFERENCES AND NOTES

(1) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P.; Gordon, E. M. Application of combinatorial Technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(2) Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P.; Gallop, M. A. Application of combinatorial Technologies to drug discovery. 2. Combinatorial organic synthesis library screening strategies, and further directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(3) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Avd. Drug. Delivery Rev.* **1997**, *23*, 3−25.

(4) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms *J. Chem. inf. Comput. Sci.* **1998**, *38*, 165−176.

(5) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs, *J. Med. Chem.* **1998**, *41*, 3325−3329.

(6) Ajay, Walters, W. P.; Murcko M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med Chem* **1998,** *41*, 3314−3324.

(7) Baldi, P.; Brunak, S. *Bioinformatics − The machine learning approach*; Cambridge MA, MIT Press: 1998.

(8) MACCS−II Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA 94577. An electronic database version of the prous science publishers journal Drug Data Report, extracted from issues starting mid-1988, contains biologically active compounds in the early stages of drug development.

(9) Available Chemicals Directory is available from MDL Informa-tion Systems Inc., San Leandro, CA, and contains specialty and bulk chemicals from commercial sources.

(10) Tripos Associates Inc., St. Louis, USA.

(11) ISIS fingerprints: SSKEYS, MDL Information System Inc., San Leandro, CA.

(12) Van Aalten, D. M. F.; Bywater, R.; Findlay, J. B. C.; Hendlich, M.; Hooft R. W. W., Vriend G. PRODRG, a program for generating molecular toplogies and unique molecular descriptors from coordinates of small molecules. *J. Comput.-Aided Mol. Design* **1996,** *10*, 255−262.

(13) Doucet, J. P.; Panaye, A.; Matthieu, G. Data processing in chemistry with neural networks: an alternative way in cases of fuzzy data or incomplete models. In *Modeling complex data for creating information;* Dubois, J. E., Gershon, N., Eds.; Springer: 1996; pp 79−88.

(14) Zimmerman, D. E.; Montelione, G. T. Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr. Opin. Struct. Biol.* **1995**, *5*, 664−673.

(15) Andrea, T. A.; Kaleyah, H. Applications of neural networks in quantitative structure−activity relationships of dihydrofolate reductase inhibition. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(16) Livingstone, D. J.; Salt, D. W. Neural networks in the search for similarity and structure−activity. In *Molecular Similarity in drug design;* Dean, P. M., Ed.; Blakie: London, 1995; pp 187−214.

(17) Leach, A. R. *Molecular modelling: principles and applications;* Addison-Wesley Longman: Harlow, 1996.

(18) Wu, C. H. Artificial neural networks for molecular sequence analysis. *Comput. Chem.* **1997**, *21*, 237−256.

(19) Qian, N.; Sejnowski, T. J. Predicting the secondary structure of globular protains using neural network models. *J. Mol. Biol*. **1988**, *202*, 865−884.

(20) MacGregor, M. J.; Flores, T. P.; Sternberg, M. J. E. Prediction of beta-turns in proteins using neural networks. *Protein Eng.* **1989**, *2*, 521−526.
(21) Rost, B.; Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.
(22) Brunak, S.; Engelbrecht, J.; Knudsen, S. 1991 Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **1991**, *220*, 49−65.
(23) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning integral representations by error propagation. In *Parallel Distributed processing: Explorations in the Microstructure of Cognition. Vol. I: Foundations*; Rumelhart, D. E., McClelland, J. L., PDP Research group, Eds.; MIT Press: Cambridge, MA, 1986; pp 318−362.
(24) Hunt, K. J.; Sbarbaro, D. Neural networks for control systems a survey, *Automatica* **1992,** *28*, 9, 1083−1112.
(25) Chen, D. Z.; Chen, Y. Q.; Hu, S. X. A pattern classification procedure integrating the multivariate statistical analysis with neural networks, *Comput. Chem.* **1997**, *21*, 2, 109−113.
(26) Mathews, B. W. Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(27) Press, W. H.; Flaneural networkery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: The art of scientific computing;* Cambridge University Press: 1986; pp 484−488.

CI0003810