Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 25 (2006) 37–45

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors Application to HIV-1 protease inhibitors

Adina-Luminiţa Milac [a], Speranţa Avram [b], Andrei-José Petrescu [a,*]

[a] Institute of Biochemistry, Splaiul Independenţei 296, Sector 6, Bucharest, Romania

[b] Faculty of Biology, Department of Biophysics and Physiology, Splaiul Independenţei 91–95, Bucharest, Romania

Received 1 April 2005; received in revised form 17 June 2005; accepted 29 September 2005
Available online 1 December 2005

Abstract

We present here a neural networks method designed to predict biological activity based on a local representation of the ligand. The compounds of the series are represented by a vector mapping for each of four substituent properties: volume, log *P*, dipole moment and a simple 'steric' parameter relating to its shape. This ligand representation was tested using neural networks on a set of 42 cyclic-urea derivatives, inhibiting HIV-1 protease. The leave-one-out cross-validation using all descriptors in the input gave a correlation factor between prediction and experiment of 0.76 for the overall set and 0.88 when three outliers were left out. To rank the significance of the four descriptors, we further tested all combinations of two and three parameters for each substituent, using two disjunctive testing sets of five inhibitors. In these sets, vectors with extreme descriptor values were used either in the training or the testing set (sets A and B, respectively). The method is a very good interpolator (set A, 95 ± 2% accuracy) but a less effective extrapolator (set B, 85 ± 2% accuracy). Generally, the combinations including the 'steric' parameter predict better than average, while those containing the volume are less effective. The best prediction, 98.8 ± 1.2%, was obtained when log *P*, the dipole and the steric parameter were used on set A. At the opposite end, the lowest ranked descriptor set was obtained when replacing log *P* with the volume, giving 92.3 ± 6.7% accuracy over the set A.

Keywords: Neural networks; QSAR; Compound library; Molecular descriptor; Biological activity prediction; HIV-1 protease inhibitors

1. Introduction

Neural networks (NN) are able to create internal models for complex input–output relationships based on learning from examples and therefore are useful in prediction.

In protein science NN were successfully used to predict secondary structure [1–3] and transmembrane segments [4], the structural class [5–8] and family [9,10], motifs such as co- and post-translational modifications [11–13], antigenic segments [14], signal sequence [15] or intracellular localization [16,17].

The NN techniques are also suited for quantitative structure–activity relationship (QSAR) applications because here a set of compounds with known activities is available for training. In contrast to simple QSAR methods based on regression analysis, where one has to priorly assume an input–output relation (e.g. linear or quadratic function), NN do not require any prior model of how input and output are connected and have the unique ability to adapt to highly complex non-linear relations [18–20]. Consequently, the essential features of NN: non-linearity, adaptivity, independence of any statistical and modelling assumptions, fault tolerance, universality and real time operation make them particularly suitable for pharmacokinetic applications, especially where extremely complex and unfamiliar responses are studied [21].

Recent examples include prediction of biological targets for chemical compounds using probabilistic NN and atom type descriptors, with 90% accuracy [22], prediction of drug resistance of HIV-1 protease mutants based on the number of drug–protein contacts, using Kohonen NN [23], selection of focussed drug libraries using feed-forward NN and 3D BCUT descriptors [24], prediction of toxicity of chemicals to aquatic species [25]. The current state-of-the-art in this field has been recently reviewed [26].

* Corresponding author. Tel.: +40 21 223 90 69; fax: +40 21 223 90 68.
E-mail address: Andrei.Petrescu@biochim.ro (A.-J. Petrescu).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.09.014

The first step in designing a NN is data pre-processing, which mainly consists in encoding the input information into an object representation so that this could be processed by the NN. This is a crucial step as the NN performance critically depends on how information is presented to the NN. An ideal encoding scheme should extract maximal information from the input data and satisfy the basic coding assumption that similar items are represented by close vectors [27]. In QSAR-like NN methods, the compounds are usually encoded by molecular descriptors—physico-chemical parameters that may be either experimental (e.g. refractive index, octanol/water partition coefficient or spectral data) or theoretical (e.g. molecular volume, weight, charge, electronic, lipophilic and steric properties).

Such a variety of parameters could generate large descriptor sets that may result either in redundancy of information—when descriptors are correlated, or chance correlations—when the dataset contains more descriptors than compounds [28]. Choosing a set of descriptors which is small enough to avoid redundancy and chance correlation, but large enough to allow an accurate representation of the ligand is therefore very important.

In this work we aim at evaluating various representations of a ligand set, focussed on substituents properties, that may be used as input in a feed-forward NN for QSAR-like applications.

The ligand set model chosen to test the method consists of 42 cyclic-urea derivatives with known inhibition constant against HIV-1 protease (PR). This system is also of practical interest and lot of data are available in the literature.

Human immunodeficiency virus type 1 (HIV-1) proteins are translated as part of the larger polyprotein precursor whose proteolytic processing during virus assembly and maturation is performed by PR [29–31]. PR is therefore an essential enzyme for HIV-1 life-cycle and a very attractive target for new antiviral drugs. The enzyme is a homodimer of 99 amino acids per chain and belong to eukaryotic aspartic protease family. The dimer has one active site region, situated at the interface between the two monomers, with one catalytic triad (Asp-Thr-Gly) from each monomer. The β-sheet configurations, which include the triplet active site Asp 25/125-Thr 26/126-Gly 27/127 are present in the major part of the enzyme (amino acids 1–85/101–185), whereas the α-helix domain is represented by the amino acids 86–99 [32,33]. PR has a high mutagenesis rate, thus being able to develop strong resistance to inhibitors [34–37]. This represents a serious problem for the anti-HIV-1 therapy. Taking into account the possible changes of the inhibitors structures, fast and precise techniques predicting the biological activity for new inhibitors are needed. In the last years the computational techniques as: molecular energy calculation [38,39], molecular docking techniques [40,41], molecular dynamics simulations [42–46] or QSAR procedures [47–54] have been useful tools for the study of the PR mutants and their inhibitors. This NN method is therefore also complementary to previous studies on interaction of HIV-1 PR inhibitors with the target enzyme.

2. Methodology

2.1. Molecular modelling of HIV-1 PR inhibitors

The set of 42 HIV-1 PR inhibitors, symmetric (benzyl, isopropyl, 4-hydroxybenzyl) cyclic-urea derivatives, was compiled from literature [55]. The criteria used for selection were: (i) the level of inhibition constants $K_i < 0.11$ nM and (ii) the variety of substituents to cyclic urea, covering as many as possible classes, e.g. methoxybenzyl, aminobenzyl, isobuthyl and hydroxybenzyl. This resulted in a highly diverse set (Table 1) in which most of the compounds have high activity. HIV-1 PR inhibitors were modelled in InsightII, starting from the cyclic-urea derivative DMP323 [56–58] complexed with HIV-1 PR (PDB code 1qbs [59]). The common cyclic urea was kept unchanged and specific substituents were added in R_1 and R_2 positions (Fig. 1A). The minimum potential energy calculations for all inhibitors were performed in Insight/Discover running conjugate-gradient method, convergence = 0.01. Electric charges of the HIV-1 PR inhibitors were loaded from InsightII dictionary applying Potentials within Force Field module.

2.2. Inhibitor parameters calculation

Each molecule is described by a vector whose elements are parameters measuring physical factors that we considered important for protein–inhibitor interaction: size (volume V), hydrophobicity (water/octanol partition coefficient log P), charge (dipole moment Δ) and shape (steric factor σ). We introduced also a steric factor to account for the orientation of structural units relative to a benzene cycle contained in R_2. This was defined as shown in Fig. 1B.

Except the steric factor which was computed only for R_2, all other parameters were computed both for R_1 and for R_2.

Molecular volume was computed with Tinker [60,61]. The hydrophobic coefficient (log P) was calculated considering the Crippen incremental value [62] using Schrodinger software. This was also used to compute the dipole moment starting form R_1/R_2 partial charges.

2.3. Neural networks

The multi-layered feed-forward NN was trained with Levenberg–Marquardt algorithm [63,64]. Due to the non-linear input–output dependency the transfer function was chosen sigmoid. All units were fully interconnected and the input-to-output information flow was feed-forward (no feed-back connections). The number of neurons in the input layer was set equal to the number of dimensions of the input vectors, while the number of neurons in the output layer was set equal to 1, i.e. the number of parameters to be predicted in this case. Based on the finding that two hidden layers with non-linear neurons are required to approximate arbitrary functions [65], we used NN having two hidden layers, each with the number of neurons taking seven

Table 1
Compound library and the scaled descriptors for each substituent; A_e: experimental activity; $A_{p\text{-}cv}$: predicted activity at cross-validation

	R_1	R_2	V_1	$\log P_1$	Δ_1	V_2	$\log P_2$	Δ_2	σ	A_e	$A_{p\text{-}cv}$
mol01	Benzyl	Benzyl	0.53	0.53	0.05	0.53	0.53	0.05	0.00	0.85	0.75
mol02	Benzyl	Methyl	0.53	0.53	0.05	**0.18**	0.25	**0.00**	0.00	**0.53**	0.64*
mol03	Benzyl	4-Isopropyl benzyl	0.53	0.53	0.05	0.80	0.85	0.02	**1.00**	0.90	0.86
mol04	Benzyl	4-(Methylthio) benzyl	0.53	0.53	0.05	0.72	0.60	0.30	**1.00**	0.85	0.85
mol05	Benzyl	Isobutyl	0.53	0.53	0.05	0.46	**1.00**	**0.00**	0.00	0.58	0.59
mol06	Benzyl	2-(Methylthio)ethyl	0.53	0.53	0.05	0.46	0.37	0.34	0.00	0.60	0.60
mol07	Benzyl	3-Indolylmethyl	0.53	0.53	0.05	0.67	0.41	0.33	0.00	0.63	0.72
mol08	Benzyl	Cyclohexylmethyl	0.53	0.53	0.05	0.66	0.83	**0.00**	0.00	0.76	0.75
mol09	Benzyl	Phenethyl	0.53	0.53	0.05	0.62	0.67	0.04	0.00	0.65	0.82*
mol10	Benzyl	2-Naphtylmethyl	0.53	0.53	0.05	0.74	0.81	0.05	0.00	0.80	0.79
mol11	Benzyl	3-Furanylmethyl	0.53	0.53	0.05	0.53	0.38	0.07	0.00	0.81	0.82
mol12	Benzyl	3-(Methylthio) benzyl	0.53	0.53	0.05	0.72	0.60	0.25	0.66	0.86	0.86
mol13	Benzyl	4-(Methylsulfonyl) benzyl	0.53	0.53	0.05	0.83	0.38	**1.00**	**1.00**	0.86	0.77
mol14	Benzyl	2-Methoxybenzyl	0.53	0.53	0.05	0.66	0.46	0.16	0.33	0.73	0.77
mol15	Benzyl	2-Hydroxybenzyl	0.53	0.53	0.05	0.55	0.49	0.16	0.33	0.75	0.74
mol16	Benzyl	3-Methoxybenzyl	0.53	0.53	0.05	0.67	0.46	0.14	0.66	0.84	0.84
mol17	Benzyl	4-Methoxybenzyl	0.53	0.53	0.05	0.67	0.46	0.18	**1.00**	0.81	0.81
mol18	Benzyl	4-Hydroxybenzyl	0.53	0.53	0.05	0.56	0.49	0.20	**1.00**	0.90	0.88
mol19	Benzyl	3-Aminobenzyl	0.53	0.53	0.05	0.60	0.33	0.25	0.66	0.86	0.84
mol20	Benzyl	3-(Dimethylamino) benzyl	0.53	0.53	0.05	0.78	0.51	0.27	0.66	0.84	0.83
mol21	Benzyl	4-Aminobenzyl	0.53	0.53	0.05	0.60	0.33	0.24	**1.00**	0.81	0.81
mol22	Benzyl	4-(Dimethylamino) benzyl	0.53	0.53	0.05	0.78	0.51	0.21	**1.00**	0.74	0.78
mol23	Benzyl	4-Pyridylmethyl	0.53	0.53	0.05	0.51	0.23	0.39	0.00	0.77	0.70
mol24	Benzyl	3-(2,5-Dimethyl pyrolyl) benzyl	0.53	0.53	0.05	**1.00**	0.93	0.48	0.66	0.68	0.72
mol25	Benzyl	3,4-(Methylenedioxy) benzyl	0.53	0.53	0.05	0.62	0.47	0.10	0.00	0.89	0.79
mol26	Cyclopropyl	Benzyl	0.41	0.52	0.01	0.53	0.53	0.05	0.00	0.88	0.75*
mol27	Cyclopropyl	Isobutyl	0.41	0.52	0.01	0.46	**1.00**	**0.00**	0.00	0.71	0.71
mol28	Cyclopropyl	2-(Methylthio)ethyl	0.41	0.52	0.01	0.46	0.37	0.34	0.00	0.56	0.56
mol29	Cyclopropyl	4-Fluorobenzyl	0.41	0.52	0.01	0.55	0.59	0.33	**1.00**	0.83	0.83
mol30	Cyclopropyl	2-methoxybenzyl	0.41	0.52	0.01	0.66	0.46	0.16	0.33	0.72	0.79
mol31	Cyclopropyl	3-Methoxybenzyl	0.41	0.52	0.01	0.67	0.46	0.14	0.66	0.91	0.82
mol32	Cyclopropyl	3-hydroxybenzyl	0.41	0.52	0.01	0.56	0.49	0.16	0.66	0.79	0.81
mol33	Cyclopropyl	4-Methoxybenzyl	0.41	0.52	0.01	0.67	0.46	0.18	**1.00**	0.86	0.86
mol34	Cyclopropyl	2-Naphthylmethyl	0.41	0.52	0.01	0.74	0.81	0.05	0.00	0.84	0.84
mol35	Hydroxybenzyl	Benzyl	0.67	0.45	0.22	0.53	0.53	0.05	0.00	0.96	0.75*
mol36	Hydroxybenzyl	2-(Methylthio)ethyl	0.67	0.45	0.22	0.46	0.37	0.34	0.00	0.54	0.72*
mol37	Hydroxybenzyl	Cyclohexylmethyl	0.67	0.45	0.22	0.66	0.83	**0.00**	0.00	0.75	0.78
mol38	Hydroxybenzyl	4-Fluorobenzyl	0.67	0.45	0.22	0.55	0.59	0.33	**1.00**	0.94	0.94
mol39	Hydroxybenzyl	3-Methoxybenzyl	0.67	0.45	0.22	0.67	0.46	0.14	0.66	**1.00**	0.97
mol40	Hydroxybenzyl	4-Pyridylmethyl	0.67	0.45	0.22	0.51	0.23	0.39	0.00	0.84	0.66*
mol41	Hydroxybenzyl	4-Methoxybenzyl	0.67	0.45	0.22	0.67	0.46	0.18	**1.00**	0.97	0.97
mol42	Hydroxybenzyl	Isobutyl	0.67	0.45	0.22	0.46	**1.00**	**0.00**	0.00	0.75	0.62*

The asterisk indicates the outlier molecules. Bold: minimal/maximal values of the descriptors.

possible values: {4, 8, 10, 15, 20, 30, 40}, resulting thus in 49 combinations. For all input data trials (different sets, different combinations of descriptors, etc.), all 49 architectures were tested and the NN giving lowest prediction error was selected.

The weights and biases of all connections were randomly initialised, then iteratively adjusted during training, according to Levenberg–Marquardt algorithm, to minimise the error function taken as the mean square error between outputs and target. The error level considered acceptable during training is

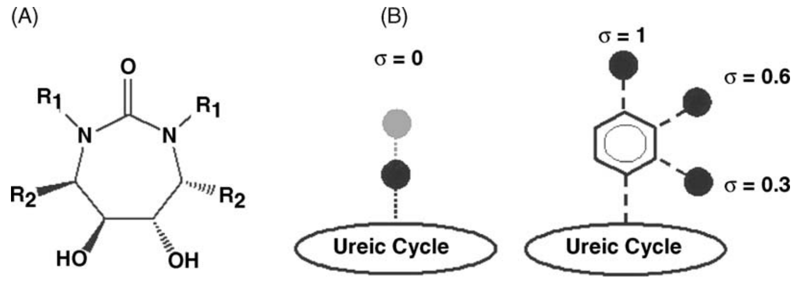

Fig. 1. (A) Chemical formulae of the inhibitors. (B) Assignment of the steric parameter, corresponding to different molecular shapes.

1%—when all the input data are learned with 99% accuracy, the training is stopped.

As the synaptic weights were randomly initialised and the starting point for error minimisation was always different, the error minimisation process might end in different local minima. Therefore, prediction could vary for the same input data and NN architecture. To assess the consistency of the results, each experiment was repeated 30 times ($n = 30$) and the results were averaged. Also due to random initialisation of the weights, in some cases the starting point for training may be located on a plateau on the multi-dimensional error surface that lead to training failures. These cases were excluded from the analysis and only successful trials were used for analysis.

2.4. Data pre-processing

2.4.1. Scaling

Prior to feeding the NN with the data, a scaling of all descriptors was performed due to two reasons. As their values span different orders of magnitude, this may lead to biases towards high-valued descriptors (e.g. volume). By scaling this artefact is avoided and all descriptors became equally important in the molecular representation. Secondly, scaling is needed as the transfer function was sigmoid, with output restricted to [0, 1].

The scaling used here was linear as previous tests [66] indicated that this procedure is optimal for this particular application.

2.4.2. Training and testing sets

The NN ability to learn and predict the biological activity was first evaluated by the "leave-one-out" cross-validation method. In this case, the molecules were described by all four parameters.

In order to test the generalization ability of the NN, we created two sets (Table 2) in which five molecules were used for testing. The composition of the sets was varied as following: set A, in which the vectors having minimum or maximum values for R_2 (most variable substituent) descriptors are used for training and the testing data contains only vectors with intermediate values; set B, in which all extreme vectors are used for testing, therefore the network has to extrapolate the data to intervals outside the training range.

Table 2
Composition of different testing sets

Set	Testing	Training	Obs
A	mol28, $A_e = 0.56$ mol07, $A_e = 0.63$ mol14, $A_e = 0.73$ mol16, $A_e = 0.84$ mol41, $A_e = 0.97$	Rest 37 molecules	Vectors with intermediate values for descriptors and output, all A_e ranges for testing
B	mol02, $A_e = 0.53$ mol42, $A_e = 0.75$ mol40, $A_e = 0.84$ mol13, $A_e = 0.86$ mol39, $A_e = 1.00$	Rest 37 molecules	Extreme vectors for testing

For both sets, the molecules were described by all possible combinations of four, three and two parameters, resulting thus in 11 input data trials.

2.5. Data post-processing

Prediction error was evaluated as the module of the difference between the experimental biological activity (A_e) and the predicted activity (A_p). Percentage error was taken as prediction error divided by the highest activity. In the case of stability tests, average prediction error (ΔA) and corresponding standard deviation (ε) were computed on the 30 repetitions as following:

$$\Delta A = |A_e - \bar{A}_p|; \quad \varepsilon = \sqrt{\frac{n\sum_{i=1}^{n} A_p^2 - \left(\sum_{i=1}^{n} A_p\right)^2}{n(n-1)}}.$$

3. Results

3.1. Assessment of the network validity and predictive ability

The ability of the network to learn the data and predict the biological activity was tested by "leave-one-out" cross-validation: 41 molecules are used for training and 1 is used for testing, this molecule being changed until the entire set is used. The results are shown in Fig. 2A. In most cases, the network is able to accurately predict the biological activity. The histogram of the prediction error (Fig. 2B) shows that prediction error lays within 2% for over 50% of the ligands and within less than 5% for two-thirds of the ligands. However, 7 out of 42 molecules (Table 1) are 'outliers' as the prediction error falls within 10–20%. Three of these molecules having activities lower than 0.7 were over-predicted ($A_p > A_e$) while four ligands having activities higher than 0.7 were under-predicted ($A_p < A_e$).

3.2. Assessment of the prediction accuracy

Using the two sets (A and B) with five testing molecules, we tested all combinations of descriptors and architectures (the best performing architecture was selected in each case). For both A and B sets, the NN is able to accurately learn all the input data as prediction reaches 99% for all molecules, regardless the composition of the set (Fig. 3).

For molecules in the testing sets results indicate that the prediction error ranges between 2% and 8% for set A, and between 10% and 16% for set B (Fig. 4). In set B, extreme vectors were not used for training, suggesting that the network ability to extrapolate data is lower than for interpolation.

3.3. The influence of molecular descriptors on prediction accuracy

For all ligand representations based on subsets of descriptors the best prediction accuracy over the 49 NN

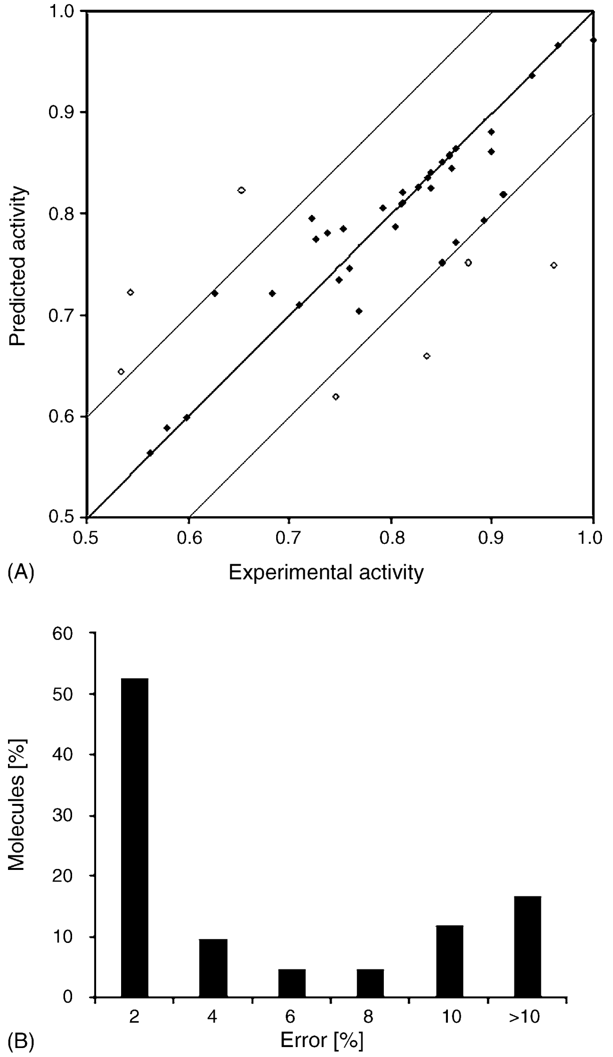

Fig. 2. Cross-validation results. (A) Dot-plot of the predicted activity vs. experimentally determined activity. Gray lines are drawn for $A_e \pm 0.1$. Molecules whose prediction error is higher than 0.1 are represented with empty circles. (B) Histogram of the relative error of the predicted activity.

architectures is presented in Fig. 5. As can be seen the method predicts better activities in the higher range, $A_e \in [0.8–1.0]$, rather than those in the lower range, $A_e \in [0.6–0.7]$. This is mainly due to the composition of the training set where low activity compounds are under-represented, 2 as compared to 22 in the $A_e \in [0.8–1.0]$ range.

Results further suggest that the largest set of descriptors (4) is redundant (with 93.5 ± 1.1% prediction accuracy), as the error obtained in some representations based on two or three descriptors is lower. The best prediction, with 98% accuracy, is obtained for the $\{\log P,\ \Delta,\ \sigma\}$ descriptor set, followed closely (97%) by the set $\{\Delta,\ \sigma\}$. By contrast adding here the volume, i.e. $\{V,\ \Delta,\ \sigma\}$ results in the lowest prediction accuracy (92%). This suggests an interesting ranking of descriptors that influence the NN sensitivity in which Δ and σ come on top while V goes to the bottom.

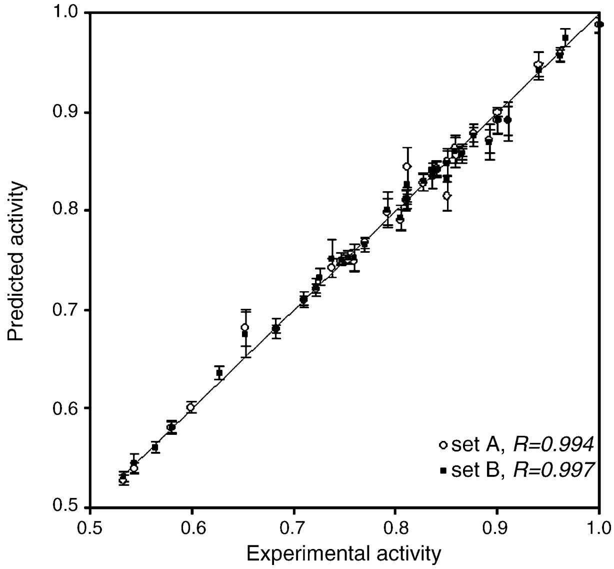

Fig. 3. Dot-plots of the predicted activity vs. experimentally determined activity for training sets A and B, when the complete set of descriptors was used as input.

3.4. The influence of network architecture on prediction accuracy

For nine most representative architectures, Fig. 6 shows the prediction error and the standard deviation over 30 runs, averaged on the testing set A when all four descriptors are used. Very similar results are obtained for all ligand representations based on descriptor subsets (data not shown).

These data indicate that network dimension correlates with prediction robustness, i.e. standard deviation (Fig. 6B), rather than prediction accuracy, i.e average prediction error (Fig. 6A). However, large networks train much slower, therefore a trade-off must be paid between robustness and computational time.

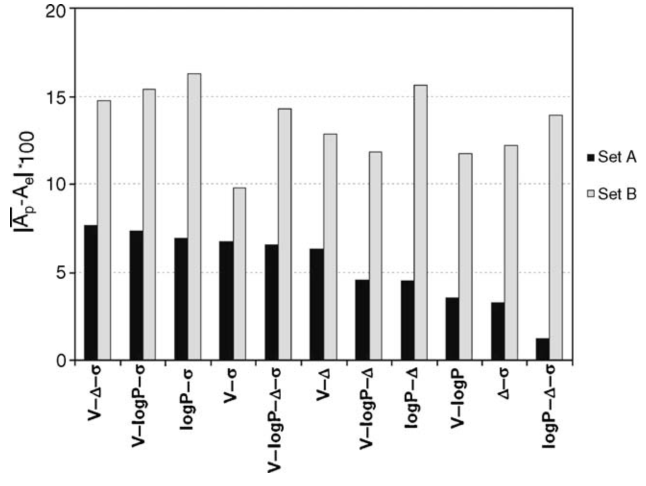

Fig. 4. Prediction error for different ligand representations using subsets of descriptors (indicated on x-axis) averaged on all molecules in testing sets A (interpolation) and B (extrapolation), indicating network poor performance at extrapolation.

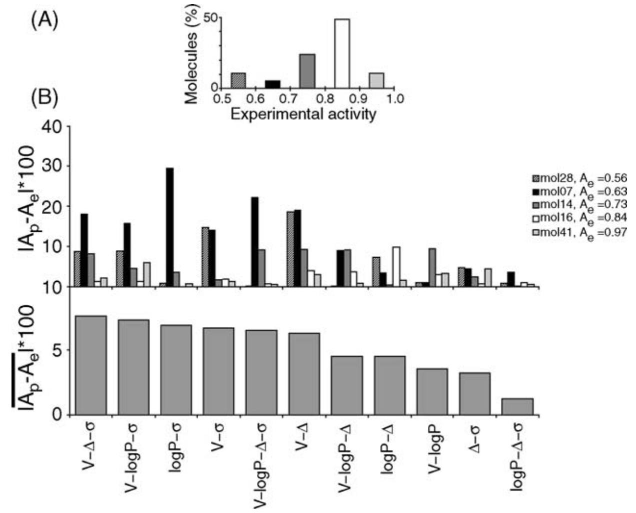

Fig. 5. (A) Histogram of the experimental activity for molecules in the training set A. (B) Prediction error, given by the difference between the predicted activity (A_p) and experimental activity (A_e), for the five molecules of the testing set A. (C) Prediction error averaged on the testing set A. All possible subsets of descriptors have been used for input encoding (*x*-axis) and are sorted from worst to best.

In addition, the analysis of the network training shows that in order to compensate for the low number of dimensions of the input vector (as in the minimal ligand representations: V–log P or Δ–σ) one has to significantly increase the number of neurons in the two hidden layers (e.g. 20 × 30 neurons for Δ–σ) resulting in a significant slowdown of training process: >160 epochs. In contrast the NN for large size vector (V–log P–Δ–σ) is small (15 × 4) but optimisation falls abruptly in false local minima, after only ~26 epochs on average. For the best ligand representation (log P–Δ–σ) on the other hand, one gets a trade-off between the number of neurons (4 × 20) and the number of training epochs (114), resulting in an optimal training process.

A detailed analysis of weights variation during training indicated that connections between the output and second hidden layer are most variable, while the innermost connections are most stable. This might be due to that these connections are in direct contact with input and output data and are most affected by update during training.

4. Discussion

In contrast with statistical methods such as multiple linear regressions, NN are able to recognize highly non-linear relationships. The flexibility of NN enables them to discover more complex relationships in experimental data, comparing with the traditional statistical models. Hence, the NN provide proper analytical alternatives to conventional techniques and interesting approaches to the QSAR and QSPR studies [67,68]. Over the past few years, the NN technique has attracted an increasing interest in drug design problems such as compound classification and multivariate calibration [69,70].

In this work, we evaluate the outcome of representing a set of ligands by mapping various local properties in a feed-forward NN method aimed at predicting their biological activity. When libraries of complex compounds need to be encoded for QSAR applications, ligands are better represented by descriptors mapping local rather than global properties. A detailed ligand representation allows more accurate modelling of protein–target interaction.

Due to the fact that cyclic-urea derivatives have two substituents R_1 and R_2, in this work we represented these molecules by a vector of variable length in which each of the two substituents are described by more than two descriptors chosen from volume (V), hydrophobicity (log P), dipole (Δ) and stericity (σ), when the substituent contains several building blocks. However, as R_1 is always a simple, mono-unit substituent, σ_1 is redundant and this reduces the maximal length of the input vector to 7.

Similar substituent-based schemes have been successfully used for input encoding in NN-based methods: QSPR applications predicting the toxicity of organophosphorus insecticides [71] and in QSAR applications on a set of benzodiazepines used in treatment of anxiety and emotional disorders [72], or on analogues of the hept inhibitors of HIV-1 reverse transcriptase [73].

Results indicate a high accuracy when the extreme input/output values are used in the training set (95%). This decreases to ~85% if extreme vectors are not comprised in the training set, which is consistent with the concept that NN are better interpolators than extrapolators when complex input–output relations are to be represented [74].

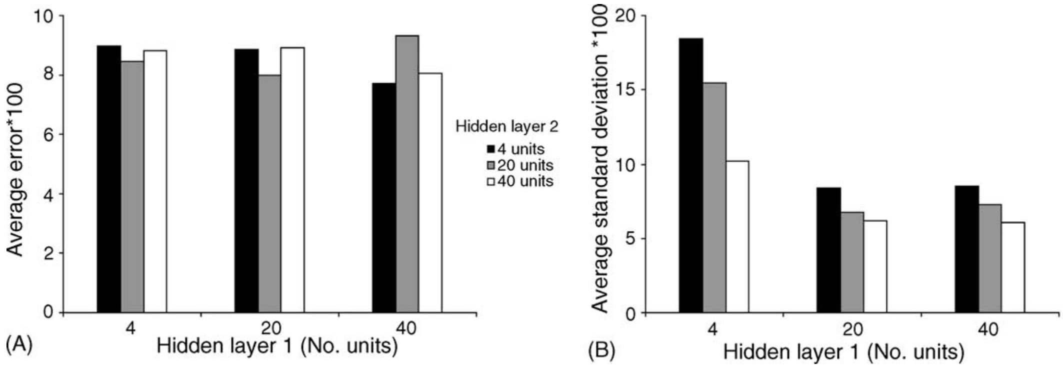

Fig. 6. (A) Average prediction error (%) and (B) standard deviation (%) for nine representative combinations of number of neurons in the first and second hidden layers, for set A, when all descriptors were used for input.

As expected, the prediction error is significantly reduced for molecules having a biological activity (output range) well represented in the training set. This is supported by the cross-validation results indicating that the three outlier molecules having activities lower than 0.7 are over-predicted ($A_p > A_e$) and the four molecules having activities higher than 0.7 are under-predicted ($A_p < A_e$), due to the fact that the composition of the entire set is biased towards molecules having high activities ($0.8 < A_e < 0.9$). Although this may result in false positives, this is less an inconvenience than excluding active compounds.

Topological descriptors emerged as a factor improving performance of NN methods in QSAR/QSPR even from early studies. Since then, a wide variety of topological descriptors have been developed. Despite their simplicity, 2D and 3D holistic topological descriptors such as the number of bonds, atom type, connectivity, molecular geometry, flexibility or surface composition increase the NN-QSAR prediction accuracy—as shown for the ligands for carbonic anhydrase isozymes [75], the angiotensin converting enzyme inhibitors [76] or for piperazyinylquinazoline analogues which exhibit PDGFR inhibition [77]. More sophisticated topological descriptors including Kappa indices, providing information about molecular shape or flexibility indices based on mass equivalence of the rotatable and rigid atoms were also very effective [78].

The steric factor (σ) introduced here to account for the substituent spatial anisotropy significantly increases the prediction accuracy, especially when this is associated with parameters describing the interaction capacity of the molecular subunit—as the dipole moment or hydrophobicity. At the opposite end, using volume as descriptor seems to have a negative impact on prediction accuracy. In addition, results indicate that substituent representations combining geometry and interaction (e.g. Δ–σ, log P–σ, V–log P) are more suitable for data encoding than combinations of same types of descriptors (e.g. V–σ).

This is in agreement with previous results indicating that proteases interact with their substrates mainly, but not exclusively, through non-covalent forces such as hydrogen bonds, ionic interactions and hydrophobic interactions [79]. For HIV-1 and HIV-2 there is a growing body of evidence suggesting that cleavage typically takes place when medium to large hydrophobic residues are present on either side of the cleavage point in the substrate [80], thus supporting the role of the hydrophobic interactions in protease function and the importance of hydrophobicity in describing the putative substrates of the enzyme. The importance of the hydrophobic effect for protease inhibition has been suggested by previous QSAR studies [52,81], this being correlated with inhibitors sliding through biological membranes. This work also indicates that an efficient HIV-1 PR inhibition could be obtained when the hydrophobic effect is correlated with a favorable electrostatic enzyme-inhibitor interaction, this possibly representing an important factor for viral resistance, especially when the mutations are placed in close vicinity of active site.

Findings similar to those presented here, suggesting the importance of steric descriptors, aside hydrophobicity, were found in a QSAR investigation of HIV-1 reverse-transcriptase inhibitors [67]. Here, the NN gave the best prediction (0.92 correlation coefficient) when using as substituent descriptors hydrophobicity and a steric factor given by the ratios between molecular length, height and width. Other combinations of geometry and physico-chemical descriptors were also useful in NN-QSAR methods predicting the activity of a set of nifedipine analogous [82] or in the design of small-molecule libraries [83] targeted for several G-protein-coupled receptor (GPCR) classes.

It is interesting to note in addition, that in contrast to the local spatial anisotropy of the ligand used here or in [73,67], the holistic topological descriptors – such as flexibility given by the degree of branching – gave less accurate results. For example, when combined with energy-related descriptors, this gave a best classification accuracy of only 85.4%, for the PTP 1B inhibitors [84] while when combined with hydrogen bonding and charge descriptors, this resulted in a best correlation coefficient of only 0.732 in the prediction of inhibitory concentration of glycine/NMDA receptor antagonist [85].

Relating to the optimal number of NN weights the debate is longstanding [86,87]. In our case, the particular NN architecture is not necessarily correlated with prediction accuracy. However, the larger networks we have tested exhibit more robust predictions (Fig. 6). This is possibly due to the fact that a large number of variables in the system would allow developing more detailed representations of the input multi-dimensional space, while few variables are able to 'cover' only general features of the input space landscape.

5. Conclusions

The results obtained by this NN method for QSAR-like applications indicate that the local mapping of ligand properties, applied here on cyclic-urea derivatives inhibiting HIV-1 protease, provide accurate results (~95%) when the extreme values are used in the training set, especially for the output ranges sufficiently well represented during training.

Results also indicate that ligand over-description results in a loss of accuracy, and here only two descriptors per substituent are sufficient in most cases for obtaining an over 95% accuracy, provided that these descriptors feed the network with balanced information (e.g. geometrical and physico-chemical).

When added, the steric factor accounting for the substituent shape, improves prediction with over 3%. Its simplicity makes it usable in other series of substituted compounds, especially when these are aromatics.

NN proves therefore accurate in predicting biological activity, provided that the input parameters are adequately chosen and enough training examples are available, being a useful tool in drug design.

References

[1] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, Proc. Natl. Acad. Sci. U.S.A. 90 (1993) 7558–7562.

[2] B. Rost, PHD: predicting one-dimensional protein structure by profile-based neural networks, Methods Enzymol. 266 (1996) 525–539.

[3] H. Kaur, G.P.S. Raghava, Prediction of β-turns in proteins from multiple alignment using neural network, Protein Sci. 12 (2003) 627–634.

[4] R. Lohmann, G. Schneider, D. Behrens, P. Wrede, A neural network model for the prediction of membrane-spanning amino acid sequences, Protein Sci. 3 (1994) 1597–1601.

[5] J. Bohr, H. Bohr, S. Brunak, R.M.J. Cotterill, H. Fredholm, B. Lautrup, S.B. Petersen, Protein structures from distance inequalities, J. Mol. Biol. 231 (1993) 861–869.

[6] B.A. Metfessel, P.N. Saurugger, D.P. Connelly, S.S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, Protein Sci. 2 (1993) 1171–1182.

[7] I. Dubchak, S.R. Holbrook, S.H. Kim, Prediction of protein folding class from amino acid composition, Proteins 16 (1993) 79–91.

[8] I. Dubchak, I. Muchnik, S.R. Holbrook, S.H. Kim, Prediction of protein folding class using global description of amino acid sequence, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 8700–8704.

[9] C.H. Wu, Gene classification artificial neural system, Methods Enzymol. 266 (1996) 71–88.

[10] C.H. Wu, S. Zhao, H.L. Chen, C.J. Lo, J. McLarty, Motif identification neural design for rapid and sensitive protein family search, Comput. Appl. Biosci. 12 (1996) 109–118.

[11] K. Julenius, A. Molgaard, R. Gupta, S. Brunak, Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites, Glycobiology 15 (2) (2004) 153–164.

[12] E.A. Berry, A.R. Dalby, Z.R. Yang, Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms, Comput. Biol. Chem. 28 (1) (2004) 75–85.

[13] L. Kiemer, J.D. Bendtsen, N. Blom, NetAcet: prediction of N-terminal acetylation sites, Bioinformatics 21 (7) (2005) 1269–1270.

[14] M. Nielsen, C. Lundegaard, P. Worning, S.L. Lauemoller, K. Lamberth, S. Buus, S. Brunak, O. Lund, Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, Protein Sci. 12 (2003) 1007–1017.

[15] B. Jagla, J. Schuchhardt, Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites, Bioinformatics 16 (3) (2000) 245–250.

[16] G. Schneider, P. Wrede, Development of artificial neural filters for pattern recognition in protein sequences, J. Mol. Evol. 36 (1993) 586–595.

[17] G. von Heijne, in: P. Wrede, G. Schneider (Eds.), Concepts in Protein Engineering and Design, Walter de Gruyter, Berlin, New York, 1994, pp. 263–279.

[18] S. Brunak, Neural Networks: Computers with Intuition, World Scientific, 1990.

[19] D.M. Skapura, Building Neural Networks, Addison-Wesley, 1995.

[20] R. Dybowski, V. Gant (Eds.), Clinical Applications of Artificial Neural Networks, Cambridge University Press, 2001.

[21] R.J. Erb, Introduction to backpropagation neural network computation, Pharm. Res. 10 (1993) 165–170.

[22] T. Niwa, Prediction of biological targets using probabilistic neural networks and atom-type descriptors, J. Med. Chem. 47 (2004) 2645–2650.

[23] S. Drăghici, R.B. Potter, Predicting HIV drug resistance with neural networks, Bioinformatics 19 (1) (2003) 98–107.

[24] M.G. Ford, W.R. Pitt, D.C. Whitley, Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks, J. Mol. Graph. Modell. 22 (2004) 467–472.

[25] K.L.E. Kaiser, The use of neural networks in QSARs for acute aquatic toxicological endpoints, J. Mol. Struct. 622 (2003) 85–95.

[26] D.A. Winkler, Neural networks as robust tools in drug lead discovery and development, Mol. Biotechnol. 27 (2) (2004) 139–168.

[27] C.H. Wu, J.W. McLarty, Neural Networks and Genome Informatics, Elsevier, 2000.

[28] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure–activity relationships, J. Med. Chem. 22 (1979) 1238–1244.

[29] R. Lapatto, T. Blundell, A. Hemmings, J. Overington, A. Wilderspin, S. Wood, J.R. Merson, P.J. Whittle, D.E. Danley, K.F. Geoghegan, X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes, Nature 342 (6247) (1989) 299–302.

[30] A. Wlodawer, M. Miller, M. Jaskolski, B.K. Sathyanarayana, E. Baldwin, I.T. Weber, L.M. Selk, L. Clawson, J. Schneider, S.B. Kent, Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease, Science 245 (4918) (1989) 616–621.

[31] A. Wlodawer, J. Vondrasek, Inhibitors of HIV-1 protease: a major success of structure-assisted drug design, Annu. Rev. Biophys. Biomol. Struct. 27 (1998) 249–284.

[32] M.A. Navia, P.M. Fitzgerald, B.M. McKeever, C.T. Leu, J.C. Heimbach, W.K. Herber, I.S. Sigal, P.L. Darke, J.P. Springer, Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1, Nature 337 (6208) (1989) 615–620.

[33] S. Piana, P. Carloni, Conformational flexibility of the catalytic Asp dyad in HIV-1 protease: an ab initio study on the free enzyme, Proteins 39 (1) (2000) 26–36.

[34] P.J. Ala, E.E. Huston, R.M. Klabe, P.K. Jadhav, P.Y. Lam, C.H. Chang, Counteracting HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with XV638 and SD146, cyclic urea amides with broad specificities, Biochemistry 37 (43) (1998) 15042–15049.

[35] R.B. Rose, C.S. Craik, R.M. Stroud, Domain flexibility in retroviral proteases: structural implications for drug resistant mutations, Biochemistry 37 (8) (1998) 2607–2621.

[36] A.C. Nair, S. Miertus, A. Tossi, D. Romeo, A computational study of the resistance of HIV-1 aspartic protease to the inhibitors ABT-538 and VX-478 and design of new analogues, Biochem. Biophys. Res. Commun. 242 (1998) 545–551.

[37] W. Markland, B.G. Rao, J.D. Parsons, J. Black, L. Zuchowski, M. Tisdale, Structural and kinetic analyses of the protease from an amprenavir-resistant human immunodeficiency virus type 1 mutant rendered resistant to saquinavir and resensitised to amprenavir, J. Virol. 74 (2000) 7636–7641.

[38] C.E. Sansom, J. Wu, I.T. Weber, Molecular mechanism analysis of inhibitor binding to HIV-1 protease, Protein Eng. 5 (2000) 659–667.

[39] C.W. Boutton, H.L. De Bondt, M.R. De Jonge, Genotype dependent QSAR for HIV-1 protease inhibition, J. Med. Chem. 48 (6) (2005) 2115–2120.

[40] M. Rarey, B. Kramer, T. Lengauer, The particle concept: placing discrete water molecules during protein-ligand docking predictions, Proteins 34 (1999) 17–28.

[41] D.J. McCarthy, J.C. Alvarez, Automated generation of MCSS-derived pharmacophoric DOCK site points for searching multiconformation databases, Proteins 51 (2003) 189–202.

[42] D.M. Ferguson, R.J. Radmer, P.A. Kollman, Determination of the relative binding free energies of peptide inhibitors to the HIV-1 protease, J. Med. Chem. 34 (1991) 2654–2659.

[43] S.W. Rick, J.W. Ericson, S.K. Burt, Reaction path and free energy calculation of the transition between alternate conformations of HIV-1 protease, Proteins 32 (1998) 7–16.

[44] L. David, R. Luo, K.G. Gilson, Comparison of generalized born and Poisson models: energetic and dynamics of HIV protease, J. Comput. Chem. 21 (2000) 295–309.

[45] M. Chun, MBO(N)D: a multibody method for long-time molecular dynamics simulations, J. Comput. Chem. 21 (2000) 159–184.

[46] E. Jenwitheesuk, R. Samudrala, Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations, BMC Struct. Biol. 3 (2003) 2–10.

[47] I. Luque, M.J. Todd, J. Gómez, N. Semo, E. Freire, Molecular basis of resistance to HIV-1 protease inhibition: a plausible hypothesis, Biochemistry 37 (1998) 5791–5797.

[48] S.P. Gupta, M.S. Babu, N. Kaw, Quantitative structure–active relationships of some HIV-protease inhibitors, J. Enzyme Inhib. 14 (1999) 109–123.

[49] U. Norinder, Refinement of catalyst hypotheses using simplex optimization, J. Comput. Aided Mol. Des. 14 (2000) 545–557.

[50] W. Schaal, A. Karlsson, G. Ahlsen, J. Lindberg, H.O. Andersson, U.H. Danielson, B. Classon, T. Unge, B. Samuelsson, J. Hulten, A. Hallberg, A. Karlen, Synthesis and comparative molecular field analysis (CoMFA) of

symmetric and nonsymmetric cyclic sulfamide HIV-1 protease inhibitors, J. Med. Chem. 44 (2001) 155–169.

[51] A. Kurup, S.B. Mekapati, R. Garg, C. Hansch, HIV-1 protease inhibitors: a comparative QSAR analysis, Curr. Med. Chem. 10 (2003) 1679–1688.

[52] S. Avram, I. Svab, C. Bologa, M.L. Flonta, Correlation between the predicted and the observed biological activity of the symmetric and nonsymmetric cyclic urea derivatives used as HIV-1 protease inhibitors. A 3D-QSAR-CoMFA method for new antiviral drug design, J. Cell. Mol. Med. 7 (2003) 287–296.

[53] C.L. Senese, A.J. Hopfinger, A simple clustering technique to improve QSAR model selection and predictivity: application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease, J. Chem. Inf. Comput. Sci. 43 (6) (2003) 2180–2193.

[54] C.L. Senese, A.J. Hopfinger, Receptor-independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease, J. Chem. Inf. Comput. Sci. 43 (4) (2003) 1297–1307.

[55] A. Nugiel, J. Seitz, Preparation and structure–activity relationship of novel P1/P1′-substituted cyclic urea-based human immunodeficiency virus type-1 protease inhibitors, J. Med. Chem. 39 (1996) 2156–2169.

[56] L. Wang, Y. Duan, P. Stouten, G.V. De Lucca, R.M. Klabe, P.A. Kollman, Does a diol cyclic urea inhibitor of HIV-1 protease bind tighter than its corresponding alcohol form? A study by free energy perturbation and continuum electrostatics calculations, J. Comput. Aided Mol. Des. 15 (2001) 145–153.

[57] R. Ishima, R. Ghirlando, J. Tözsér, A.M. Gronenborn, D.A. Torchia, J.M. Louis, Folded monomer of HIV-1 protease, J. Biol. Chem. 276 (2001) 49110–49116.

[58] J. Boisbouvier, A. Bax, Long-range magnetization transfer between uncoupled nuclei by dipole–dipole cross-correlated relaxation: a precise probe of beta-sheet geometry in proteins, J. Am. Chem. Soc. 124 (2002) 11038–11045.

[59] P.Y. Lam, Y. Ru, P.K. Jadhav, P.E. Aldrich, G.V. De Lucca, C.J. Eyermann, C.H. Chang, G. Emmett, E.R. Holler, W.F. Daneker, L. Li, P.N. Confalone, R.J. McHugh, Q. Han, R. Li, J.A. Markwalder, S.P. Seitz, T.R. Sharpe, L.T. Bacheler, M.M. Rayner, R.M. Klabe, L. Shum, D.L. Winslow, D.M. Kornhauser, C.N. Hodge, Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure–activity relationship, and molecular recognition of cyclic ureas, J. Med. Chem. 30 (1996) 3514–3525.

[60] R.V. Pappu, R.K. Hart, J.W. Ponder, Analysis and application of potential energy smoothing for global optimization, J. Phys. Chem. B 102 (1998) 9725–9742.

[61] P. Ren, J.W. Ponder, Polarizable atomic multipole water model for molecular mechanics simulation, J. Phys. Chem. B 107 (2003) 5933–5947.

[62] A.K. Ghose, A. Pritchett, G.M. Crippen, Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships. III: Modelling hydrophobic interactions, J. Comput. Chem. 9 (1988) 80–90.

[63] K. Levenberg, A method for the solution of certain problems in least squares, Q. Appl. Math. 2 (1944) 164–168.

[64] D. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, SIAM J. Appl. Math. 11 (1963) 431–441.

[65] G. Schneider, P. Wrede, Artificial neural networks for computer-based molecular design, Prog. Biophys. Mol. Biol. 70 (1998) 175–222.

[66] A.L. Milac, S. Avram, A.J. Petrescu, A new neural networks method for predicting biological activity of chemical compounds, Rom. J. Biochem. 40 (1–2) (2003) 35–45.

[67] D.T. Manallak, D.D. Ellis, D.J. Livingstone, Analysis of linear and nonlinear QSAR data using neural networks, J. Med. Chem. 37 (1994) 3758–3767.

[68] M. Shamsipur, B. Hemmateenejad, M. Akhond, Multicomponent acid–base titration using principal component-artificial neural network calibration, Anal. Chim. Acta 461 (2002) 147–153.

[69] T. Khanna, Foundations of Neural Networks, Addison-Wesley, New York, 1991.

[70] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533.

[71] M. Zahouily, A. Rhihil, H. Bazoui, S. Sebti, D. Zakarya, Structure–toxicity relationships study of a series of organophosphorus insecticides, J. Mol. Model. 8 (5) (2002) 168–172.

[72] S.S. So, M. Karplus, Genetic neural networks for quantitative structure–activity relationships: improvements and application of benzodiazepine affinity for benzodiazepine/GABAA receptors, J. Med. Chem. 39 (26) (1996) 5246–5256.

[73] H. Bazoui, M. Zahouily, S. Boulajaaj, S. Sebti, D. Zakarya, QSAR for anti-HIV activity of HEPT derivatives, SAR QSAR Environ. Res. 13 (6) (2002) 567–577.

[74] S. Snider, C. Allende Prieto, T. Von Hippel, T.C. Beers, C. Sneden, Y. Qu, S. Rossi, Astrophys. J. 562 (2001) 528–548.

[75] B.E. Mattioni, P.C. Jurs, Development of quantitative structure–activity relationship and classification models for a set of carbonic anhydrase inhibitors, J. Chem. Inf. Comput. Sci. 42 (1) (2002) 94–102.

[76] J.J. Sutherland, L.A. O'Brien, D.F. Weaver, A comparison of methods for modeling quantitative structure–activity relationships, J. Med. Chem. 47 (22) (2004) 5541–5554.

[77] R. Guha, P.C. Jurs, Development of linear, ensemble, and nonlinear models for the prediction and interpretation of the biological activity of a set of PDGFR inhibitors, J. Chem. Inf. Comput. Sci. 44 (6) (2004) 2179–2189.

[78] A.K. Madan, S. Gupta, M. Singh, Superpendentic index: a novel highly discriminating topological descriptor for predicting biological activity, J. Chem. Inf. Comput. Sci. 39 (1999) 272–277.

[79] A. Narayanan, X. Wu, Z.R. Yang, Mining viral protease data to extract cleavage knowledge, Bioinformatics 18 (2002) S5–S13.

[80] S.C. Pettit, J. Simsic, D.D. Loeb, L. Everitt, C.A. Hutchinsin III, R. Swanstrom, Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid, J. Biol. Chem. 266 (1991) 14539–14547.

[81] R. Garg, B. Bhhatarai, A mechanistic study of 3-aminoindazole cyclic urea HIV-1 protease inhibitors using comparative QSAR, Bioorg. Med. Chem. 12 (22) (2004) 5819–5831.

[82] B. Hemmateenejad, M. Akhond, R. Miri, M. Shamsipur, Genetic algorithm applied to the selection of factors in principal component-artificial neural networks: application to QSAR study of calcium channel antagonist activity of 1,4-dihydropyridines (nifedipine analogous), J. Chem. Inf. Comput. Sci. 43 (4) (2003) 1328–1334.

[83] K.V. Balakin, S.A. Lang, A.V. Skorenko, S.E. Tkachenko, A.A. Ivashchenko, N.P. Savchuk, Structure-based versus property-based approaches in the design of G-protein-coupled receptor-targeted libraries, J. Chem. Inf. Comput. Sci. 43 (5) (2003) 1553–1562.

[84] S.J. Patankar, P.C. Jurs, Classification of inhibitors of protein tyrosine phosphatase 1B using molecular structure based descriptors, J. Chem. Inf. Comput. Sci. 43 (3) (2003) 885–899.

[85] S.J. Patankar, P.C. Jurs, Prediction of glycine/NMDA receptor antagonist inhibition from molecular structure, J. Chem. Inf. Comput. Sci. 42 (5) (2002) 1053–1068.

[86] J. Zupan, J. Gasteiger, Neural Networks for Chemists. An Introduction, VCH, Weinheim, Germany, 1993.

[87] J.V. Turner, D.J. Cutler, I. Spence, D.J. Maddalena, Selective descriptor pruning for QSAR/QSPR studies using artificial neural networks, J. Comput. Chem. 24 (7) (2003) 891–897.