

A genetic algorithm optimized fuzzy neural network analysis of the affinity of inhibitors for HIV-1 protease

Levente Fabry-Asztalos,^{a,*} Răzvan Andonie,^b Catharine J. Collar,^a
Sarah Abdul-Wahid^b and Nicholas Salim^a

^aDepartment of Chemistry, Central Washington University, Ellensburg, WA 98926, USA

^bDepartment of Computer Science, Central Washington University, Ellensburg, WA 98926, USA

Received 12 October 2007; revised 22 December 2007; accepted 28 December 2007

Available online 1 January 2008

Abstract—A fuzzy neural network (FNN) was trained on a dataset of 177 HIV-1 protease ligands with experimentally measured IC₅₀ values. A set of descriptors was selected to build nonlinear quantitative structure–activity relationships. A genetic algorithm (GA) was implemented to optimize the architecture of the fuzzy neural network used to predict biological activity of HIV-1 protease inhibitors. Evolutionary methods were used to apply feature selection (FS) to this model. Results obtained on an external test set of 21 molecules, with and without feature selection, were compared. Applying feature selection to the GA-FNN resulted in a more accurate prediction of biological activity. Fuzzy IF/THEN rules were extracted from the optimized FNN. In the future the developed models are expected to be useful in the rational design of novel enzyme inhibitors for HIV-1 protease.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

For the discovery of novel enzyme inhibitors and their optimization we consider computational modeling as a valuable aid. Software for molecular modeling, neural networks, and statistical analysis are applied to data sets to generate predictive models for biological activities.^{1–6} Neural networks were used to predict biological activities of HIV-1 reverse transcriptase and vesicular monoamine transporter-2 inhibitors, drug resistance, lipophilicity, aqueous solubility, intestinal absorption, and site of protease cleavage.^{7–13} Computational models can assist medicinal chemists in more quickly identifying and optimizing novel lead compounds, possibly improving the overall efficiency of the drug discovery process.

Sybyl and BioLoom are commonly used to collect compounds and their corresponding molecular descriptors.^{14–18} Training and testing molecules can be obtained from a single databank, multiple databanks, or from a compilation of several previous studies.^{17–19} For supervised training, the dataset must also contain

experimental data, such as values for IC₅₀ (inhibitory concentration at 50%) or K_i (the dissociation constant of the inhibitor).

Neural networks (NN) are frequently designed using algorithms that require supervised training to obtain predictions.^{20–22} Fuzzy logic algorithms can enhance the NN by relating structure to function with IF/THEN rules. This network is referred to as a fuzzy neural network (FNN).¹⁸ Evolutionary methods can be used to optimize the architecture of the FNN, by determining the best error tolerance and the optimal subset of descriptors (feature selection).^{23–25} Data mining techniques can be employed to extract rules, rank importance of molecular descriptors, and determine the contribution of molecular descriptors to biological predictions.^{21,26–30}

Statistics are used to assess prediction accuracy. Common measures for analyzing multiple predictions are root mean squared error (RMSE), symmetrical mean absolute percentage error (sMAPE), Pearson's correlation coefficient (*r*), and the coefficient of determination (*r*²).^{18,31,32}

In this study we implement a GA optimized FNN to produce two models for inhibitors of HIV-1 protease.

Keywords: HIV-1 protease inhibitors; Fuzzy neural network; IF/THEN rules; QSAR; Genetic algorithm; Feature selection.

* Corresponding author. Tel.: +1 509 963 2887; fax: +1 509 963 1050; e-mail: fabryl@cwu.edu

The first uses all 35 molecular descriptors (GA-FNN). The second applies feature selection using evolutionary methods (FS-GA-FNN). Fuzzy IF/THEN rules are extracted from the FS-GA-FNN.

2. Methods

2.1. Inhibitory compounds for the genetic algorithm optimized fuzzy neural network

Experimental datasets for HIV-1 protease were collected.^{33–37} Redundant molecules, molecules missing experimentally determined biological values, and those with conflicting biological activities were removed from the datasets leaving a total of 198 HIV-1 protease inhibitors. These 198 molecules were divided into a set for training and a set for testing. A set with 21 molecules was selected to be the test set. The IC₅₀ values for the test set molecules range from 1.4 nM to 3000 nM with a mean value of 157.33 nM. The remaining 177 molecules were used for training and cross-validation. The IC₅₀ values for the training and cross-validation set molecules range from 0.28 nM to 11,800 nM with a mean value of 683.16 nM. After all phases of training and optimization were completed using only the training set, the FNN's accuracy was evaluated using the test set.

All molecules were drawn into Sybyl where they were minimized, aligned, and docked within the protease active site. Tripos force field was used and minimization was performed using Gasteiger–Marsili charges.^{38,39} When aligning molecules there was no algorithm used, they were aligned by common atoms. The molecules were aligned prior to docking to make sure that the starting structures were in a similar conformation. To dock compounds into the active site the docking option within SYBYL base was used. Structures were minimized to convergence using a gradient of 0.01 kcal/(mol Å) and a maximum of 1000 iterations.

2.2. Molecular descriptors

We designed software to extract 35 molecular descriptors from the files generated by Sybyl and BioLoom. These descriptors are listed in Figure 1. Selection of molecular descriptors was based on accessibility and contribution to molecular entity. The summation of atoms and bonds (D01 and D11) along with the number of each type of atom (D02–D08), the numbers of each type of bond (D12–D16), and functional group (D16 and D17) provided atomic descriptors for each inhibitory structure. It has been shown that there is a correlation between molecular weight and biological affinity of inhibitory compounds. Thus, molecular weights (D17), the number of hydrogen bond donors (D31), the number of hydrogen bond acceptors (D32), and ClogP (D33) were also included. Index of hydrogen deficiency (D10) was incorporated to describe the degrees of unsaturation for each molecule. Molecular volume (D09), calculated molar refractivity (D34), and the number of valence electrons (D35) were calculated to represent the overall three-dimensional space that molecules occu-

No.	Molecular Descriptors
D01	Total Number of Atoms
D02	Number of Carbon Atoms
D03	Number of Chlorine Atoms
D04	Number of Fluorine Atoms
D05	Number of Hydrogen Atoms
D06	Number of Nitrogen Atoms
D07	Number of Oxygen Atoms
D08	Number of Sulfur Atoms
D09	Molecular Volume
D10	Index of Hydrogen Deficiency
D11	Total Number of Bonds
D12	Number of Single Bonds
D13	Number of Double Bonds
D14	Number of Triple Bonds
D15	Number of Aromatic Bonds
D16	Number of Amide Bonds
D17	Molecular Weight
D18	Total Charge
D19	Bond Stretching Energy
D20	Angle Bending Energy
D21	Torsional Energy
D22	Out of Plane Bending Energy
D23	One to Four Van Der Waals Energy
D24	Van Der Waals Energy
D25	One to Four Electrostatic Energy
D26	Electrostatic Energy
D27	Total Energy
D28	Van Der Waals Electrostatic Pairs
D29	One to Four Van Der Waals Electrostatic Pairs
D30	Scaled Van Der Waals Electrostatic Pairs
D31	Hydrogen Bond Donors
D32	Hydrogen Bond Acceptors
D33	Calculated LogP
D34	Calculated Molar Refractivity
D35	Number of Valence Electrons

Figure 1. Molecular descriptors used for the training and testing of the GA-FNN.

py. Total charge (D18), total energy (D27), molecular energy descriptors (D19–D26), and summations of molecular interactions (D28–D30) established the conformational molecular relationships for each inhibitory structure. Descriptors D01 through D30 were generated from SYBYL, D19 through D30 resulting from energy calculations of the minimized structures using the same force field and charges as the ones used for minimization. Descriptors D31 through D35 were generated using BioLoom.

2.3. Normalization

Descriptor values were normalized to produce standardized ranges between 0 and 1:

$$\text{Normalized value} = (\text{Actual} - \text{Min}) / (\text{Max} - \text{Min})$$

In the above equation, the actual value is the molecular descriptor value obtained from Sybyl or BioLoom. The minimum (Min) is the smallest value of the molecular descriptor in the dataset, while the maximum (Max) is the largest value of the molecular descriptor in the dataset (Table 1). Before training and cross-validation, the range of all molecules (test, training/validation, and no-

Table 1. Minimum and maximum values of molecular descriptors and the corresponding ranges for the five fuzzy membership groups for the data set compounds

Molecular descriptors	Min	Max	Low [min, low]	Med-low [low, med-low]	Med [med-low, med]	Med-high [med, med-high]	High [med-high, high]
Total number of atoms	33	125	51.4	69.8	88.2	106.6	125
Number of carbon atoms	17	49	23.4	29.8	36.2	42.6	49
Number of chlorine atoms	0	1	0.2	0.4	0.6	0.8	1
Number of fluorine atoms	0	3	0.6	1.2	1.8	2.4	3
Number of hydrogen atoms	12	61	21.8	31.6	41.4	51.2	61
Number of nitrogen atoms	0	8	1.6	3.2	4.8	6.4	8
Number of oxygen atoms	3	10	4.4	5.8	7.2	8.6	10
Number of sulfur atoms	0	3	0.6	1.2	1.8	2.4	3
Molecular volume	226.19	715.27	324.01	421.82	519.64	617.45	715.27
Index of hydrogen deficiency	7	26	10.8	14.6	18.4	22.2	26
Total number of bonds	35	129	53.8	72.6	91.4	110.2	129
Number of single bonds	20	95	35	50	65	80	95
Number of double bonds	2	8	3.2	4.4	5.6	6.8	8
Number of triple bonds	0	1	0.2	0.4	0.6	0.8	1
Number of aromatic bonds	6	30	10.8	15.6	20.4	25.2	30
Number of amide bonds	0	5	1	2	3	4	5
Molecular weight	296.34	900.09	417.09	537.84	658.59	779.34	900.09
Total charge	−1	468.80	92.96	186.92	280.88	374.84	468.80
Bond stretching energy	0.33	17.48	3.76	7.19	10.62	14.05	17.48
Angle bending energy	2.50	104.03	22.81	43.11	63.42	83.72	104.03
Tortional energy	3.34	27.08	8.09	12.83	17.58	22.33	27.08
Out of plane bending energy	0	0.50	0.10	0.20	0.30	0.40	0.50
One to four Van Der Waals energy	1.17	12.01	3.34	5.50	7.67	9.84	12.01
Van Der Waals energy	−19.74	−0.59	−15.91	−12.08	−8.25	−4.42	−0.59
One to four electrostatic energy	−13.83	227.77	34.49	82.81	131.1	179.45	227.77
Electrostatic energy	−24.62	141.31	8.57	41.75	74.94	108.13	141.31
Total energy	3.96	478.64	98.89	193.83	288.76	383.70	478.64
Van Der Waals electrostatic pairs	368	6685	1631.4	2894.8	4158.2	5421.6	6685
One to four Van Der Waals electrostatic pairs	69	320	119.2	169.4	219.6	269.8	320
Scaled Van Der Waals electrostatic pairs	2	62	14	26	38	50	62
Hydrogen bond donors	1	8	2.4	3.8	5.2	6.6	8
Hydrogen bond acceptors	2	15	4.6	7.2	9.8	12.4	15
Calculated log <i>P</i>	1.08	9.96	2.86	4.63	6.41	8.18	9.96
Calculated molar refractivity	8.46	24.65	11.70	14.94	18.17	21.41	24.65
Number of valence electrons	104	344	152	200	248	296	344

vel potential inhibitor sets) was determined, so that all normalized values fell within the [0, 1] range.

2.4. Statistics

The statistics used to evaluate predictive ability are symmetric mean absolute percentage error (sMAPE), root mean squared error (RMSE), and Pearson product-moment coefficient of correlation (Pearson's *r*). For sMAPE and RMSE, lower values are optimal, while *r*, in this case, is best when closest to 1.¹⁸

3. The genetic algorithm optimized fuzzy neural network

The FNN was implemented according to a modification of the Min–Max Fuzzy Inference Network (MMFIN) described by Cai and Kwan.⁴⁰ This modification consists of replacing the *d*–*y* error (where *d* is the target value and *y* is the computed value) of the MMFIN learning algorithm, with the symmetric relative error, (*d*–*y*)/(*d* + *y*). This incorporates the fact that small differences of low IC₅₀ values are chemically more signifi-

cant than the same amount of difference of high IC₅₀ values. Note that all *d* and *y* are positive in our FNN. The network consists of three layers: the vector of input values; the hidden layer; and the output layer, consisting of one neuron which is the predicted IC₅₀ value for the molecule. Each node of the hidden layer represents a prototype of the input vectors. The number of hidden layer neurons is determined by the training algorithm, with the maximum possible being the number of ligands used for training.

The input vector layer and hidden layer are connected by fuzzy membership functions. The fuzzy membership value of each input molecular descriptor is calculated, in relation to previously learned input vectors, and stored in a matrix. The first dimension represents the descriptors and the second dimension represents the hidden layer neurons. For each molecule passing through the network, the membership value of each descriptor in each of the previously learned prototypes is determined.

The hidden layer and output layer are connected by a weight matrix. The first dimension of this weight matrix

represents the hidden layer neurons and the second dimension represents the output neuron. The output of the hidden layer is the minimum (fuzzy AND) of the membership function input. This hidden layer output is then multiplied by the weight matrix. The resultant matrix serves as input to the third (output) layer of the network. The output of the third layer of the network is the maximum (fuzzy OR) of this input, and is the computed IC_{50} value of the input vector of descriptors.

During the training process, the final output is compared to the target value, which is the known biological affinity value of the training ligand. If the output differs from this value by more than the established error tolerance, the membership functions and weight matrix are adjusted and this ligand must be 're-learned'. If it is not possible to re-learn this molecule so that it falls within the ranges of the already established prototypes (hidden layer neurons), a new neuron representing this prototype is added to the hidden layer. Training continues until all training molecules are learned and produce output values within the acceptable error tolerance.

The FNN architecture is determined by the error tolerance; the challenge is how to find the best value for this parameter. The goal of optimizing the structure of the FNN with a specified set of descriptors is to produce the most compact network possible, which still maintains a high prediction and generalization ability. These optimization criteria are measured by a fitness function which incorporates both the number of hidden nodes and the predictive ability of the FNN.

The objective of the GA optimization is to find the optimum balance between the extremes of a highly compact network, which is unable to predict accurately, and a very loose, overfitted network, which is not able to generalize. We use the same GA in two different instances: first to determine the optimal subset of features and second to determine the optimal error tolerance for the FNN.

To incorporate evolutionary methods into feature selection, we used the idea introduced by Siedlecki and Sklansky.⁴¹ They implemented a GA to find an optimal binary vector, where each bit is associated with a feature. If the i th bit of the vector equals 1, then the i th feature is allowed to participate in prediction; if the bit is 0, then the corresponding feature does not participate. Each resulting subset of features is evaluated in accordance to the fitness function f on a set of training data, using the FNN and cross-validation.

The evolutionary methods used incorporate two types of chromosomes: (1) An fChromosome in which each gene is either a 0 or a 1; this chromosome defines the subset of features with which to train the FNN. (2) An eChromosome, in which each gene is an integer; all genes are in the range [0,9], except the first gene which is Refs. 8,9 This chromosome is converted to a float, which is used as the error tolerance to train the FNN. The resultant error tolerance therefore falls

within the range [0.8,1.0] The minimum value of 0.8 was selected following numerous sequential optimizations; it was noted that when the error tolerance is less than 0.8, the number of hidden nodes is too high, resulting in overfitting.

The GA incorporates both crossover and mutation operators. The population consists of 70 individuals. Using the roulette wheel method, 34 individuals are selected for crossover. For each pair, the point of crossover is randomly determined and the genes beyond this point are exchanged. From the resultant population, seven individuals are selected for mutation using the roulette wheel method. For each individual, the point of mutation is selected randomly. In the case of FS, in which each gene is either a 0 or 1; the gene is mutated by flipping the current value to its opposite value. In the case of error tolerance optimization, in which each gene is an integer, the new gene value is selected randomly within the correct range.

4. Stages of the genetic algorithm optimized fuzzy neural network

Our method can be described by the following four stages:

(1) *FS optimization*: An initial population of fChromosomes is instantiated. To obtain an estimate of the fitness of each member of this population, 200 eChromosomes are randomly generated for each fChromosome. Tenfold cross-validation is used to train and validate the FNN with each eChromosome for each fChromosome, using only the training set of molecules. An eChromosome encodes the error tolerance that is used to train the FNN. The FNN architecture is governed by this error tolerance and so it is important to optimize this value. The optimal error tolerance (eChromosome) for each set of descriptors (fChromosome) was determined in order to achieve the optimal architecture for this set of descriptors. During cross-validation, the sMAPE, RMSE, and r are calculated, and the average number of hidden nodes (aveM) is determined. The fitness of each fChromosome using each eChromosome, is evaluated according to the following formula:

$$f = (1/(sMAPE + 2 \cdot aveM)) + 0.01 \cdot r \quad (1)$$

The fitness value ultimately assigned to each fChromosome is the fittest value obtained from evaluating all 200 eChromosomes with this particular subset of descriptors. This represents a coarse-grained estimate of the fitness of each fChromosome. Random search only is used to identify the best eChromosome for each fChromosome; no evolutionary methods such as mutation or crossover are applied at this stage to the eChromosomes. The evolutionary methods (survival of the fittest, cross-over, and mutation) are applied to the generations of fChromosomes. The final result of this stage is an fChromosome which represents the optimized set of features.

(2) *Error tolerance optimization*: The subset of features used is defined by the fittest fChromosome obtained from the FS optimization. An initial population of eChromosomes is generated. It is seeded by the eChromosome obtained during the estimate of the fittest eChromosome during the FS optimization.

The leave-one-out (LOO) cross-validation method is used to calculate the sMAPE, RMSE and r . The average number of hidden nodes (aveM) in the trained FNN is determined. The fitness value of each eChromosome is calculated according to Eq. (1). The FNN trained with the fittest fChromosome and the fittest eChromosome is the result of this stage. For the GA-FNN, this step only is applied to the complete set of 35 descriptors.

(3) *Test set evaluation*: The next step is to evaluate the ability of the FNN obtained from Step 2 to make accurate predictions. This is accomplished using the test set of 21 molecules. Each molecule in this set is passed through both the GA-FNN and the FS-GA-FNN to predict its IC_{50} value. The sMAPE, RMSE, and Pearson's r are calculated.

(4) *Rule extraction*: The final step is to extract the fuzzy IF/THEN rules from the FNN produced by Step 2. Each hidden node corresponds to a learned prototype as well as to a fuzzy IF/THEN rule. This rule is accessible at the end of the learning phase. IF/THEN rules encapsulate the explicit knowledge of the FNN. The number of rules corresponds to the number of nodes within the optimized FNN. Rules describe the relationship between the descriptor values and the predicted IC_{50} . The normalized range of $[0, 1]$ is divided into five equal groups (low, low-medium, medium, medium-high, and high). The ranges without normalization that correspond to the five equal groups are shown in Table 1 above. Categories for IC_{50} were divided into low (0–20 nM), low-medium (20–50 nM), medium (50–100 nM), medium-high (100–500 nM), and high (>500 nM). Information generated by the IF/THEN

rules is very valuable when evaluating known and potential inhibitory compounds.

5. Results and discussion

Figure 2 illustrates the total population fitness as evolutionary pressure is applied to the feature defining chromosomes. The population's overall fitness improves with successive generations. The final average fitness for all members in the population was 0.01858043. The global best individual, with a fitness value of 0.020195693, was obtained after 17 generations.

Further optimization of the error tolerance for this global best individual was performed. As result, the following statistics were obtained for the training set with LOO cross-validation. The predictive ability of the FS-GA-FNN and the GA-FNN are displayed in Table 2. The sMAPE and RMSE statistics for the FS-GA-FNN with the reduced set of 15 descriptors are lower than those obtained from the GA-FNN which used all 35 descriptors. The r value for the FS-GA-FNN is closer to 1 than the r value for the GA-FNN. This indicates that the FS-GA-FNN outperforms the GA-FNN.

All three statistics improve when the reduced subset of features is used. This would indicate that some of the 35 descriptors are not relevant for predicting the biological activity of the ligand. The global best individual obtained with the evolutionary feature selection defined a

Table 2. Statistics on the training set for the FS-GA-FNN and GA-FNN

	FS-GA-FNN	GA-FNN
sMAPE	83.4	98.4
RMSE	890.4	1289.9
Pearson's r	0.82	0.68
Number of hidden nodes	18	27
Number of descriptors	15	35

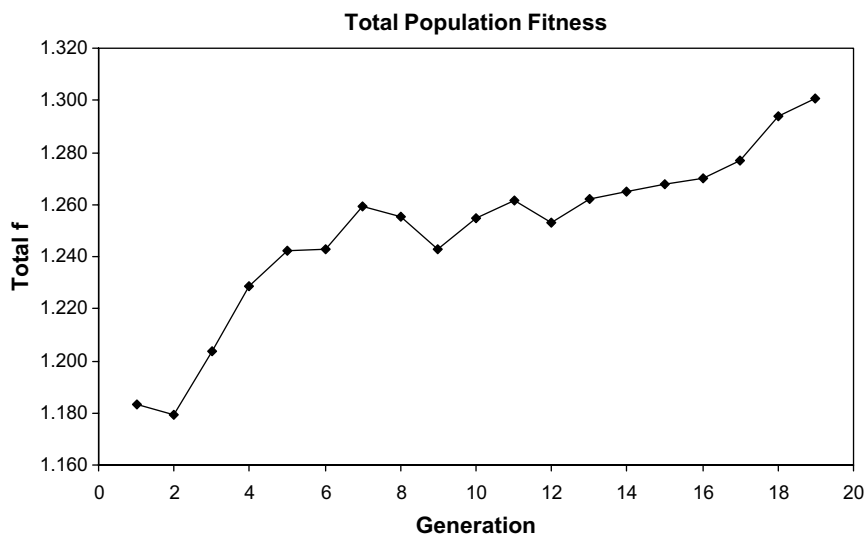


Figure 2. The total population fitness.

subset of 15 features, listed in Figure 3, and was used to make predictions on the test set of molecules.

This improvement in statistical results using the reduced set of features is also seen in the results for the test set molecules (Table 3). With the reduced subset of features a more compact and optimized neural network structure is obtained, as evidenced by the fact that only 18 hidden nodes are obtained, compared to 27 when all 35 features are used. This more compact architecture provides an improved generalization capability to the network.

No.	Molecular Descriptors
D04	Number of Fluorine Atoms
D05	Number of Hydrogen Atoms
D07	Number of Oxygen Atoms
D09	Molecular Volume
D10	Index of Hydrogen Deficiency
D11	Total Number of Bonds
D19	Bond Stretching Energy
D22	Out of Plane Bending Energy
D24	Van Der Waals Energy
D25	One to Four Electrostatic Energy
D26	Electrostatic Energy
D27	Total Energy
D28	Van Der Waals Electrostatic Pairs
D29	One to Four Van Der Waals Electrostatic Pairs
D34	Calculated Molar Refractivity

Figure 3. Molecular descriptors selected.

Table 3. Statistics on the test set for the FS-GA-FNN and GA-FNN

	FS-GA-FNN	GA-FNN
sMAPE	108.2	166.3
RMSE	513.4	722.6
Pearson's <i>r</i>	0.9818	−0.2781

Graphical representation of theoretical (predicted) versus experimental (actual) biological affinity values for the test set molecules is shown in Figure 4.

We have also predicted IC₅₀ values for 26 newly designed compounds. The results and chemical structures are shown in Tables 4–6. These novel compounds

Table 4. Novel inhibitory structures with their corresponding predicted IC₅₀ values using FS-GA-FNN and GA-FNN

No.	R1	R2	Predicted IC ₅₀ (nM)	
			FS-GA-FNN	GA-FNN
1	OCH ₃	Cbz	17.24	34.29
2	Cbz	OCH ₃	23.02	76.13
3	Cbz	COH	20.94	34.29
4	Cbz	NH ₂	5.33	41.60
5	Cbz	OSO ₂ Ph-(4-CN)	1925.43	16.48
6	OCH ₃	COH	240.78	0.45
7	OCH ₃	NH ₂	231.43	0.40
8	OCH ₃	OSO ₂ Ph-(4-CN)	23.88	12.65
9	Ph-OH	COH	5.64	41.35
10	Furan	NH ₂	648.82	0.57
11	Furan	OSO ₂ Ph-(4-CN)	9.58	0.86
12	Thiophene	NH ₂	651.99	0.51
13	Thiophene	OSO ₂ Ph-(4-CN)	9.58	0.87
14	4-Methylthiazole	NH ₂	73.63	0.50
15	4-Methylthiazole	OSO ₂ Ph-(4-CN)	9.58	0.87

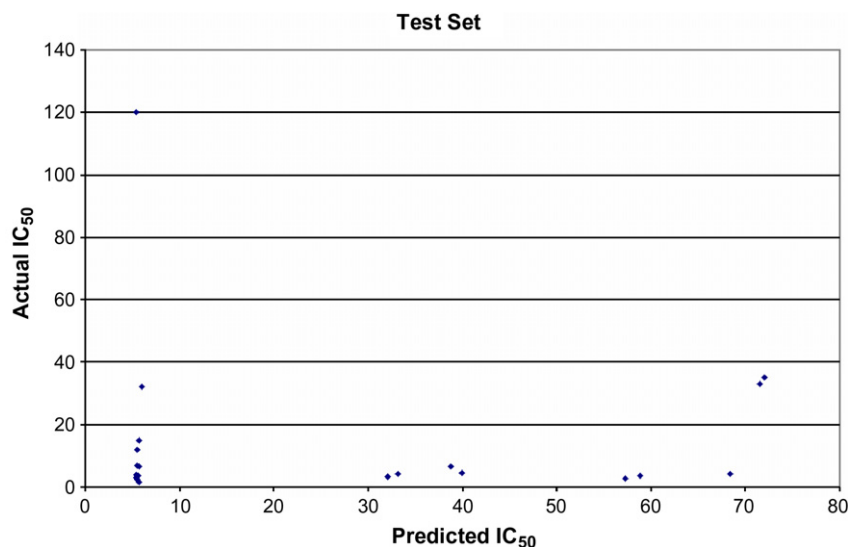
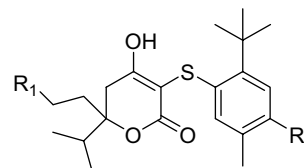
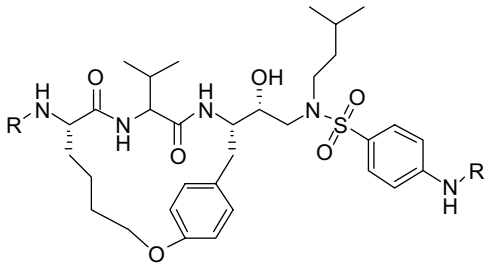
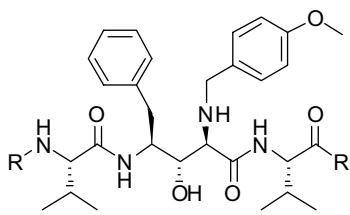


Figure 4. Graphical representation of cross-validation results using FS-GA-FNN prediction, showing predicted versus experimental biological affinity values (in nM) for the test set molecules (leaving out the one molecule with the experimental IC₅₀ value of 3000 nM).

Table 5. Novel inhibitory structures with their corresponding predicted IC_{50} values using FS-GA-FNN and GA-FNN


No.	R	Predicted IC_{50} (nM)	
		FS-GA-FNN	GA-FNN
16	OCH ₃	1783.13	1783.13
17	COH	1656.63	1656.63
18	Cbz	1285.71	1285.71
19	Furan	1891.57	1000.00
20	Thiophene	1873.49	4.67
21	4-Methylthiazole	1861.45	1.75

Table 6. Novel inhibitory structures with their corresponding predicted IC_{50} values using FS-GA-FNN and GA-FNN


No.	R	Predicted IC_{50} (nM)	
		FS-GA-FNN	GA-FNN
22	OCH ₃	2171.71	1800.00
23	COH	25.26	1500.00
24	Cbz	1714.29	1714.29
25	Furan	2329.22	1500.00
26	Thiophene	2448.98	37.25

were designed based on known inhibitory structures with experimentally determined IC_{50} and TI values by leaving their core structure intact and replacing their side chains with new ones coming from other inhibitors. Based on these predictions, we have initiated synthesis on the most promising molecules from Table 4. Since both of the methods used (FS-GA-FNN and GA-FNN) predicted high inhibitory activity for some of the novel compounds shown in Table 4 (1–4, 6–9, 11, and 13–15) they are good candidates for synthesis and further drug discovery.

In a previous study, we have compared results obtained with our fuzzy neural networks with results obtained using Multiple Linear Regression (MLR).¹⁸ Generally, our neural network approach performed better. However, in the MLR study we used only 151 known structures with experimentally determined IC_{50} values for training and cross-validation while in the current study 177.

The IF/THEN rules extracted from the FS-GA-FNN are shown in Figures 5–7. For example Rule 1 should be read as: if D04, D07, D10, D19, D22, and D25–D27 are low-medium and D05, D09, D11, D24, D28, D29, and D34 are medium then the IC_{50} value is evaluated as low (see Figures 3 and 5). The corresponding descriptor ranges for the five groups for the training set compounds is shown in Table 1 above.

When evaluating these rules we look for common trends that medicinal chemists can use to design new enzyme inhibitors. These rules are specific to the enzyme system and theoretical model used in our current study. The designed novel structures 1–4, 6–9, 11, and 13–15 shown in Table 4 are predicted to have good biological affinity because their physicochemical properties match the desired characteristics described by the IF/THEN rules. Further studies with other enzyme systems and theoretical models may give us a better understanding of really how specific or general these rules are.

RULES EVALUATING TO LOW IC_{50} :		RULE 1				RULE 2				RULE 3				RULE 4				RULE 11				RULE 12				RULE 13			
IF		L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H
Number of Fluorine Atoms																													
Number of Hydrogen Atoms																													
Number of Oxygen Atoms																													
Molecular Volume																													
Index of Hydrogen Deficiency																													
Total Number of Bonds																													
Bond Stretching Energy																													
Out Of Plane Bending Energy																													
Van Der Waals Energy																													
One To Four Electrostatic Energy																													
Electrostatic Energy																													
Total Energy																													
Van Der Waals Electrostatic Pairs																													
One To Four Van Der Waals Electrostatic Pairs																													
Calculated Molar Refractivity																													
THEN, IC_{50} :																													

Figure 5. IF/THEN rules for low IC_{50} predictions.

RULES EVALUATING TO MEDIUM LOW, MEDIUM, AND MEDIUM HIGH IC50:				RULE 5				RULE 6				RULE 14				RULE 15				RULE 7				RULE 16			
IF				L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H	L	ML	MM	H
Number of Fluorine Atoms																											
Number of Hydrogen Atoms																											
Number of Oxygen Atoms																											
Molecular Volume																											
Index of Hydrogen Deficiency																											
Total Number of Bonds																											
Bond Stretching Energy																											
Out Of Plane Bending Energy																											
Van Der Waals Energy																											
One To Four Electrostatic Energy																											
Electrostatic Energy																											
Total Energy																											
Van Der Waals Electrostatic Pairs																											
One To Four Van Der Waals Electrostatic Pairs																											
Calculated Molar Refractivity																											
THEN, IC50:																											

Figure 6. IF/THEN rules for medium low, medium, and medium high IC₅₀ predictions.

RULES EVALUATING TO HIGH IC50:				RULE 8				RULE 9				RULE 10				RULE 17				RULE 18							
IF				L	M	L	M	M	M	H	H	L	M	L	M	M	M	H	H	L	M	L	M	M	M	H	H
Number of Fluorine Atoms																											
Number of Hydrogen Atoms																											
Number of Oxygen Atoms																											
Molecular Volume																											
Index of Hydrogen Deficiency																											
Total Number of Bonds																											
Bond Stretching Energy																											
Out Of Plane Bending Energy																											
Van Der Waals Energy																											
One To Four Electrostatic Energy																											
Electrostatic Energy																											
Total Energy																											
Van Der Waals Electrostatic Pairs																											
One To Four Van Der Waals Electrostatic Pairs																											
Calculated Molar Refractivity																											
THEN, IC50:																											

Figure 7. IF/THEN rules for high IC₅₀ predictions.

6. Conclusion

When feature selection was applied to the GA-FNN, more accurate prediction of biological activity was obtained. From this, one can conclude that many of the 35 descriptors are not relevant for predicting IC₅₀ values. This information, especially when combined with a careful examination of the fuzzy IF/THEN rules extracted, can provide valuable insight into which features of the ligand are crucial to obtaining a highly inhibitory molecule. This method has the potential to greatly assist medicinal chemists in the design of lead compounds for HIV-1 protease, as well as for other therapeutically important enzymes.

Acknowledgments

The authors thank the Faculty Research Fund, the Science Honors Program, and the Office of Graduate Studies and Research at Central Washington University for partial support.

References and notes

- Nair, A. C.; Jayatilke, P.; Wang, X.; Miertus, S.; Welsh, W. J. *J. Med. Chem.* **2002**, *45*, 973–983.
- Jorgensen, W. L. *Science* **2004**, *303*, 1813–1818.
- Beger, R. D.; Buzatu, D. A.; Wilkes, J. G.; Lay, J. O., Jr. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1360–1366.
- Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 718–726.
- Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218–1227.
- Andonie, R.; Fabry-Asztalos, L.; Abdul-Wahid, S.; Col-lar, C.; Salim, N. *Proc. IEEE Int. Conf. Neural Netw.*, **2006**, 7495–7502.
- Tetko, I. V.; Tanchuk, V. Y.; Luik, A. I. *Proceedings of the Seventh Annual IEEE Symposium on Computer Based Medical Systems*, **1994**, 311–316.
- Draghici, S.; Potter, R. B. *Bioinformatics* **2003**, *19*, 98–107.
- Weekes, D.; Fogel, G. B. *BioSystems* **2003**, *72*, 149–158.
- Zheng, F.; Zheng, G.; Deaciuc, A. G.; Zhan, C. G.; Dwoskin, L. P.; Crooks, P. A. *Bioorg. Med. Chem.* **2007**, *15*, 2975–2992.
- Tetko, I. V.; Tanchuk, V. Y. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136–1145.
- Niwa, T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113–119.
- Yang, Z. R.; Thomson, R. *IEEE Trans. Neural Netw.* **2005**, *16*, 263–274.
- SYBYL Molecular Modeling Software*, 7.2 ed.; Tripos Inc.: St. Louis, MO, 2006.
- BioLoom, Version 2004*; BioByte Corp.: Claremont, CA **2004**.

16. Ang, W. H.; Scopelliti, R.; Bussy, F.; Juillerat-Jeanneret, L.; Dyson, P. J. *J. Med. Chem.* **2005**, *48*, 8060–8069.
17. Peng, Y.; Keenan, S. M.; Zhang, Q.; Kholodovych, V.; Welsh, W. J. *J. Med. Chem.* **2005**, *48*, 1620–1629.
18. Andonie, R.; Fabry-Asztalos, L.; Collar, C. J.; Abdul-Wahid, S.; Salim, N. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, **2005**, 113–120.
19. Faulon, J. L.; Visco, D. P., Jr.; Pophale, R. S. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.
20. Yao, X. J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257–1266.
21. Niwa, T. *J. Med. Chem.* **2004**, *47*, 2645–2650.
22. Loukas, Y. L. *J. Med. Chem.* **2001**, *44*, 2772–2783.
23. Tetko, I. V.; Solov'ev, V. P.; Antonov, A. V.; Yao, X.; Doucet, J. P.; Fan, B.; Hoonakker, F.; Fourches, D.; Jost, P.; Lachiche, N.; Varnek, A. *J. Chem. Inf. Model* **2006**, *46*, 808–819.
24. Mosier, P. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1460–1470.
25. Mosier, P. D.; Counterman, A. E.; Jurs, P. C.; Clemmer, D. E. *Anal. Chem.* **2002**, *74*, 1360–1370.
26. Cecchini, M.; Kolb, P.; Majeux, N.; Cafisch, A. *J. Comput. Chem.* **2004**, *25*, 412–422.
27. Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.
28. Tetko, I. V.; Bruneau, P. *J. Pharm. Sci.* **2004**, *93*, 3103–3110.
29. Tetko, I. V.; Poda, G. I. *J. Med. Chem.* **2004**, *47*, 5601–5604.
30. Tetko, I. V.; Villa, A. E.; Aksenova, T. I.; Zielinski, W. L.; Brower, J.; Collantes, E. R.; Welsh, W. J. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 660–668.
31. Senese, C. L.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1297–1307.
32. Giordanetto, F.; Cotesta, S.; Catana, C.; Trosset, J. Y.; Vulpetti, A.; Stouten, R. F. W.; Kroemer, R. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 882–893.
33. Boyer, F. E.; Prasad, J. V. N. V.; Domagala, J. M.; Ellsworth, E. L.; Gaida, C.; Hagen, S. E.; Markovski, L. J.; Tait, B. D.; Lunney, E. A.; Palovski, A.; Ferguson, D.; Graham, N.; Holler, T.; Hupe, D.; Nouhan, C.; Tummino, P. J.; Urumov, A.; Zeikus, G.; Gracheck, S. J.; Sanders, J. M.; VanderRoest, S.; Brodfuehrer, J.; Iyer, K.; Sinz, M.; Gulnik, S. V.; Erickson, J. W. *J. Med. Chem.* **2000**, *43*, 843–858.
34. Hagen, S. E.; Domagala, J.; Gajda, C.; Lovdahl, M.; Tait, B. D.; Wise, E.; Holler, T.; Hupe, D.; Nouhan, C.; Urumov, A.; Zeikus, G.; Zeikus, E.; Lunney, E. A.; Pavlovsky, A.; Gracheck, S. J.; Saunders, J.; VanderRoest, S.; Brodfuehrer, J. *J. Med. Chem.* **2001**, *44*, 2319–2332.
35. Hagen, S. E.; Prasad, J. V. N. V.; Boyer, F. E.; Domagala, J. M.; Ellsworth, E. L.; Gajda, C.; Hamilton, H. W.; Markoski, L. J.; Steinbaugh, B. A.; Tait, B. D.; Lunney, E. A.; Tummino, P. J.; Ferguson, D.; Hupe, D.; Nouhan, C.; Gracheck, S. J.; Saunders, J. M.; VanderRoest, S. *J. Med. Chem.* **1997**, *40*, 3707–3711.
36. Scholz, D.; Billich, A.; Charpiot, B.; Ettmayer, P.; Lehr, P.; Rosenwirth, B.; Schreiner, E.; Gstach, H. *J. Med. Chem.* **1994**, *37*, 3079–3089.
37. Tait, B. D.; Hagen, S.; Domagala, J.; Ellsworth, E. L.; Gajda, C.; Hamilton, H. W.; Prasad, J. V. N. V.; Ferguson, D.; Graham, N.; Hupe, D.; Nouhan, C.; Tummino, P. J.; Humblet, C.; Lunney, E. A.; Pavlovsky, A.; Rubin, J.; Gracheck, S. J.; Baldwin, E. T.; Bhat, T. N.; Erickson, J. W.; Gulnik, S. V.; Liu, B. *J. Med. Chem.* **1997**, *40*, 3781–3792.
38. Klon, A. E.; Glick, M.; Davies, J. W. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216–2224.
39. Peterson, S. D.; Schaal, W.; Karlen, A. *J. Chem. Inf. Model* **2006**, *46*, 355–364.
40. Cai, L. Y.; Kwan, H. K. *IEEE Trans. Syst. Man Cybern. B* **1998**, *28*, 334–347.
41. Siedlecki, W.; Sklansky, J. *Pattern Recognit. Lett.* **1989**, *10*, 335–347.