A hybrid decision trees–adaptive neuro–fuzzy inference system in prediction of anti-HIV molecules

ARTICLE *in* EXPERT SYSTEMS WITH APPLICATIONS · MAY 2011

Impact Factor: 2.24 · DOI: 10.1016/j.eswa.2010.11.011 · Source: DBLP

2 AUTHORS:

Mohamed Kissi
Université Chouaib Doukkali
21 PUBLICATIONS **32** CITATIONS
SEE PROFILE

Mohammed Ramdani
Université Hassan II de Casablanca
74 PUBLICATIONS **89** CITATIONS
SEE PROFILE

Expert Systems with Applications 38 (2011) 6376–6380

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A hybrid decision trees-adaptive neuro-fuzzy inference system in prediction of anti-HIV molecules

Mohamed Kissi [a,*], Mohammed Ramdani [b]

[a] Equipe MMID, Département de Mathématiques et Informatique, Faculté des Sciences, B.P. 20, 24000 El Jadida, Morocco
[b] Département d'informatique (LIM@II), Faculté des Sciences et Techniques, B.P. 146, 20650 Mohammedia, Morocco

ARTICLE INFO

Keywords:
Fuzzy inference system
Decision trees
QSAR
Anti-HIV

ABSTRACT

Several works quantitative structure–activity relationships (QSAR) of anti-human immunodeficiency virus (HIV) molecules were studied by different statistical methods and non-linear models. But few studies have used the heuristic methods. In this paper, a hybrid decision trees (DT) and adaptive neuro-fuzzy inference system (ANFIS) is used for the prediction of inhibitory activity of anti-VIH molecules. DT algorithm is utilized to select the most important variables in QSAR modeling and then these variables were used as inputs of ANFIS to predict the anti-HIV activity. The model's predictions were compared with other methods and the results indicated that the proposed models in this work are superior over the others.

1. Introduction

Acquired immuno deficiency syndrome (AIDS) is still an illness with lethal evolution. Still, there is no complete effective chemotherapy. The disease is caused by a retrovirus called the human immunodeficiency virus (Redfield, Blattner, & Gallo, 1999). The virus annihilates the organism's natural capacity for self-defense: it compromises the immune system and creates a favorable environment to develop various infective diseases (opportunistic infections) with increased lethal potential. The same mechanism is involved in the development of various neoplasias. The nervous system represents also a target of HIV with severe neurological and psychiatrical consequences. Once the viral replication has reached an explosive rate, the chances to control the disease are minimal (Carne, 1987; Smith & Spittle, 1987). These data are strong motivations for research especially in the field of structure–activity relationships among different classes of viral inhibitors and also for knowledge of HIV and its consequences at the molecular level.

QSAR searches information relating chemical structure to biological and other activity by developing a QSAR model. By using such an approach one could predict the activities of newly designed compounds before a decision is being made whether these compounds should be really synthesized and tested. Building a QSAR model begins with calculating theoretical parameters for the compounds involved. The experimental information associated with biological properties is taken as dependent variables in developing a model. Several descriptors could be generated in QSAR studies, but only some of them are statistically significant in terms of correlation with biological activity for a particular analysis. The challenge, therefore, is to select the subset of descriptors that describe the most critical structural and physicochemical features associated with inhibitory activity. Effective descriptor or variable selection is an integral part of the QSAR modeling process (Ghafourian & Cronin, 2005). Obtaining a good quality QSAR model depends on many factors, such as the quality of biological data, the choice of descriptors and statistical methods. There have been many variable selection methods, the mostly used ones are stepwise regression, neural network, fuzzy logic and genetic algorithms (Cho & Hermsmeier, 2002; Yasri & Hartsough, 2001). In this study, a variable selection technique was proposed based, this method is namely Decision Trees (DT) (Quinlan, 1986). As a novel computational approach, DT algorithms have attracted attention of the researchers in many fields (Kissi, Ramdani, Bouchon-Meunier, & Zakarya, 2008; Marsala, 1998; Ramdani, 1994; Ramdani, Kissi, & Bouchon-Meunier, 2004).

For developing a reliable QSAR model, adaptive neuro-fuzzy inference system (ANFIS) was used. The synergism of fuzzy logic systems and neural network has produced a functional system capable of learning high-level thinking and reasoning (Zadeh, 1977). It is an improved tool for determining the behavior of imprecisely defined complex systems. The purpose of a neuro-fuzzy system is to apply neural learning techniques to identify the parameters and/or structure of neuro-fuzzy systems. The fuzzy systems have fascinated the growing attention and interest in bioinformatics applications, decision making studies, pattern recognition, and data analysis (Kissi, Ramdani, Tollabi, & Zakarya,

* Corresponding author. Tel.: +212 5 23 34 30 03; fax: +212 5 23 34 21 87.
E-mail address: kissim@gmail.com (M. Kissi).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.11.011

2004; Marsala, Ramdani, Tollabi, & Zakarya, 2000; Ros, Pintore, & Chretien, 2002).

In the present work, we have examined DT algorithm for variable selection and ANFIS for model developing in QSAR analysis of inhibitory anti-HIV activity. The results of DT-ANFIS were compared to those of stepwise multiple linear regression (MLR) (Bazoui, Zahouily, Boulaajaj, Sebti, & Zakarya, 2002), artificial neural networks (ANN) (Douali, Cherqaoui, & Villemin, 2004; Jalali-Heravi & Parastar, 2000) and 3D QSAR partial least-squares (PLS) (Ravichandran, Prashantha Kumar, Sankar, & Agrawal, 2009). It has been demonstrated that the DT is a useful tool for variable selection comparable to the stepwise MLR, ANN and 3D QSAR-PLS.

2. Theoretical routines

2.1. Decision trees

A decision tree is a natural structure of knowledge. Each node in such a tree is associated with a test on values of an attribute, each edge from the node is labeled with a particular value of the attribute, and each leaf of the tree is associated with a value of the class. Each path in the tree is associated with a production rule, where premises are composed by the tests associated with nodes and the conclusion of the rule is composed by the class associated with the leaf of the path

2.1.1. Induction learning

A decision tree can be constructed from a set of examples by inductive learning. Inductive learning is a process to generalize knowledge from the set of examples. The set of examples is called the training set. Each example is described by the values of a set of attributes and the values of an attribute class. The decision tree is built from its root to its leaves. The training set is successively split by means of questions on the value of a chosen attribute. Let E be a training set composed by a set of examples ei. Each ei is defined by a value $ei(Aj)$ for each attribute Aj of a set of attributes A, and the value $ei(C)$ of the attribute class C. The algorithm to build decision trees is the following (Quinlan, 1986):

- Select an attribute Aj, with mj values $\{vj1, \ldots, vjmj\}$.
- Split the training set E by means of the selected attribute Aj. Each value vjl is associated with a subset Ejl of E:

$$E = \bigcup_{l=1,\ldots,mi} E_l \quad \text{with } E_l = \{e_i \in E | e_i(A_j) = v_{jl}\}.$$

- Each value vjl of the attribute labels an edge from the current node. This edge leads to the corresponding subset Ejl. In the case of the continuous attributes the training examples are sorted on their values, a threshold t is selected allowing to split the training set into subsets $Ej1 = \{ei \in E | ei(Aj) > t\}$ and $Ej2 = \{ei \in E | ei(Aj) < t\}$. In this case the edges from the current node are labeled with $Aj \leqslant t$ and $Aj > t$.
- Check if the stopping criterion is fulfilled by each subset Ejl. A leaf is created from a subset that meets this criterion. This leaf is associated with the majority class of this subset. Start again in step 1 with all the subsets that do not meets the stopping criterion. The selected attribute, the threshold for continuous attribute and a stopping criterion can be based on a measure of discrimination.

2.1.2. Measure of discrimination

A measure of discrimination enables us to measure the power of discrimination of an attribute Aj relatively to the class C. It evaluates to which extent the values of this attribute are linked with each value of the class. In the process of the construction of decision trees, the choice of the best attribute by means of a measure of discrimination is a heuristic that enables us to optimize the tree being built. This heuristic should minimize the size of the tree.

A measure of discrimination that we choose is a Shannon entropy:

$$H(C/A_j) = -\sum_{l=1}^{mj} p(V_{jl}) \sum_{k}^{K} p(C_k/V_{jl}) \log(p(C_k/V_{jl})),$$

where $V_{jl} = \{e_j \in E / e_i(A_j) = v_{jl}\}$ and $C_k = \{e_i \in E / e_i(C) = c_k\}$

2.1.3. Stopping criterion

The stopping criterion can be based on:

- A threshold of the measure of discrimination: the smaller value of the measure of discrimination corresponding to a training set homogeneous with regard to the class.
- The number of examples in E, if it is too small to justify a split of E.

2.2. Fuzzy inference systems

Fuzzy inference systems (FIS) may be used as tools for approximating ill-defined nonlinear functions. They can import qualitative aspects of human knowledge and reasoning processes by data sets without employing precise quantitative analysis using the following five functional components as shown in Fig. 1:

- A rule base containing a number of fuzzy if–then rules.
- A data base defining the membership functions of fuzzy sets.
- A decision-making unit as the inference engine.
- A fuzzification interface which transforms crisp inputs to linguistic variables.
- A defuzzification interface converting fuzzy outputs to crisp outputs.

ANFIS is an architecture, which is functionally equivalent to a Takagi-Sugeno fuzzy rule base, whose parameters are tuned by a learning algorithm using input–output data available (Takagi & Sugeno, 1985). Assume a simple Takagi–Sugeno FIS with two inputs x and y and one output f and a rule base with two fuzzy if–then rules as follows:

- **Rule 1**. If x is A_1 and y is B_1 then $f_1 = p_1x + q_1y + r_1$,
- **Rule 2**. If x is A_2 and y is B_2 then $f_2 = p_2x + q_2y + r_2$,

where A_1, A_2 and B_1, B_2 are, respectively, fuzzy sets of input premise variables x and y; and p_1, q_1, r_1 and p_2, q_2, r_2 are parameters of the consequent or output variables. The general structure of ANFIS is presented in Fig. 2 wherein square nodes are fixed nodes and circle nodes are adaptive nodes whose parameters are changed during the training process. A hybrid learning algorithm is employed in ANFIS by which the parameters of membership functions of input variables in antecedent part of fuzzy rules are optimized using a steepest descent algorithm while the linear parameters of the output variable in the consequent part are optimized using a least

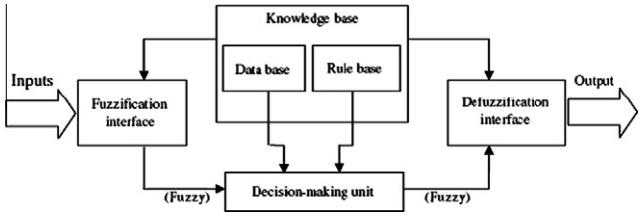

Fig. 1. FIS architecture.

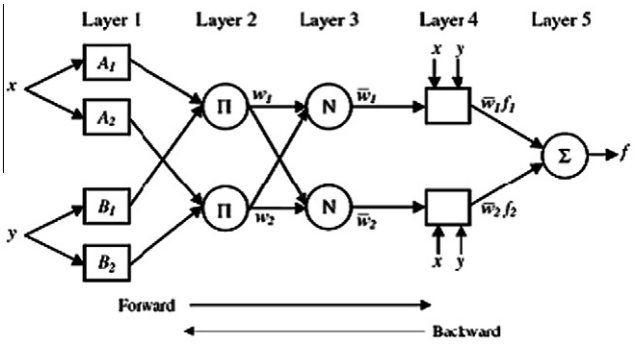

Fig. 2. ANFIS model structure.

square method. The final output of the given network with two inputs and one output in terms of the above parameters can be calculated as follows (Jang, 1993):

$$f = \frac{w_1}{w_1+w_2}f_1 + \frac{w_2}{w_1+w_2}f_2 = \bar{w}_1 f_1 + \bar{w}_2 f_2$$
$$= \bar{w}_1(p_1x + q_1y + r_1) + \bar{w}_2(p_2x + q_2y + r_2)$$
$$= (\bar{w}_1x)p_1 + (\bar{w}_1y)q_1 + \bar{w}_1r_1) + (\bar{w}_2x)p_2 + (\bar{w}_2y)q_2 + \bar{w}_2r_2),$$

$$w_i = \mu_{Ai}(x)\mu_{Bi}(y) \quad i = 1,2,$$
$$\mu_{Ai}(x) = \exp\left[-\left(\frac{x-c_i}{a_i}\right)^2\right],$$
$$\mu_{Bi}(y) = \exp\left[-\left(\frac{y-d_i}{e_i}\right)^2\right] \quad i = 1,2,$$

where w_i is called the firing strength of rule i, $\mu_{Ai}(x)$ and $\mu_{Bi}(y)$ are, respectively the membership degrees of x and y in A_i and B_i, c_i and d_i are, respectively, the mean values of the Gaussian membership functions defined for fuzzy sets A_i and B_i and a_i and e_i are, respectively, the standard deviations of the membership functions.

3. Experimental

3.1. Data set

The data set consists of 79 anti-HIV molecules together with their inhibitory activities and is taken from the articles published by Tanaka et al. (1992) and Garg, Gupta, Gao, Babu, and Debnath (1999). The anti-HIV activity of the compounds has been expressed by the compound's ability to protect cells against the cytopathic effect of the virus. This activity is the concentration required to 50% effect has been measured and expressed as pIC_{50} (50% peptide inhibitory concentration).

3.2. Molecular descriptors

A main step in every QSAR study is choosing and calculating the structural descriptors as numerical encoded parameters representing the chemical structures. In this work, each molecule was described by seven descriptors, which are given by Garg et al. (1999). These descriptors characterize hydrophobic, steric, geographic and electronic aspects. The variable molecules used are:

- *LogP*: logarithm of partition coefficient between water and octanol of molecule.
- X_1: connectivity index 4th order type path.
- X_2: connectivity index of 6th order of string.
- S: molecule surface.
- B_1, B_2: parameters of molecule substituents.
- *RM*: molecular substituting reactivity of molecule.

4. Results and discussion

4.1. Hybrid DT-ANFIS model

As it will be mentioned in the previous section, the number of the features that was used for estimating pIC_{50} was seven. The dimension of the feature vector was reduced to three with the DT algorithm which was described in Section 2.1. The hybrid learning algorithm has been used in ANFIS structure because it is highly efficient in training. Fig. 3 shows the flowchart of DT and ANFIS used in this study.

4.2. Selection of pertinent variables with decision trees

For the purpose of the selection of the most important variables, DT was applied using seven descriptors as independent variables and observed pIC_{50} as a dependent variable. Each example described by values for these descriptors is associated with a class (pIC_{50}). Take an example of the database studied, the description is associated with pIC_{50} (class = 1), to obtain a learning example (Table 1).

In this work the activity pIC_{50} is continued, the training with decision trees generates many rules that the chemist expert cannot interpret. To solve this problem the activity was divided into two classes: weak activity and strong activity. Once all examples are described by these seven different descriptors, the class is coded by a binary variable indicating a weak (class 1) or strong (class 2) presence of pIC_{50}, and database is processed to develop a decision trees which binds structure of molecules and their activity. The obtained tree is given with the following rules:

- **Rule 1:** if $X_1 \leqslant 0.27$ then class 1 (80%).
- **Rule 2:** if $X_1 > 0.27$ and $X_2 \leqslant 0.028$ and $S > 1.78$ then class 2 (86%).

Using the rules found, it concludes that the best descriptors selected are X_1 to identify the class 1 and X_1, X_2 and S to identify the class 2.

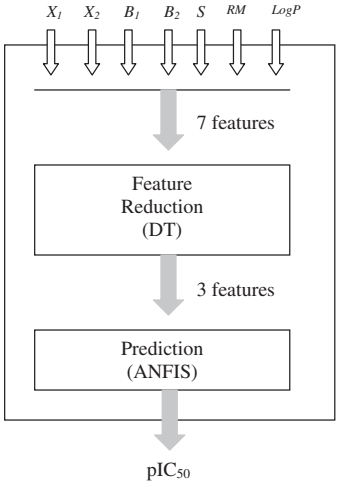

Fig. 3. Stages of modeling with DT-ANFIS.

Table 1
Description and class of example molecule.

X_1	B_1	X_2	B_2	*LogP*	*RM*	S	pIC_{50}
0.26	1	0.042	1.52	1.87	0.57	0.65	1

4.3. Hybrid DT-ANFIS model

Decision trees are applied to identify variables responsible for the activity subdivided on only two classes. In this part, we consider the numeric score activity (Fig. 4) and the attributes selected by the decision trees. An ANFIS has been developed to predict pIC_{50} activity. Three machining parameters namely X_1, X_2 and S were taken as input features (Fig. 3). Building an ANFIS model with a minimum number of fuzzy rules would help escape from a well-known problem, i.e. curse of dimensionality. To cluster the training data set, several methods like hard c-mean, fuzzy c-mean or subtractive clustering could be used. The subtractive clustering (Chiu, 1994) method works based on the potential of any data points in a feature space to be cluster centers. The points with high potential value are first selected as candidates to be the cluster centers. The points whose distance from a cluster center is less than a pre-specified value (cluster radius) are subtracted then and the potential values are iteratively updated. The rule extraction method first uses the subtractive clustering function to determine the number of rules and antecedent membership functions and then uses linear least squares estimation to determine each rule's consequent equations. This function returns a ANFIS structure that contains a set of fuzzy rules to cover the feature space.

The system generated seven rules:

- If X_1 is near to 0.27 and X_2 is near to 0.042 and S is near to 1.66 then pIC_{50} is near to 5.24.
- If X_1 is near to 0.29 and X_2 is near to 0.038 and S is near to 1. 66 then pIC_{50} is near to 8.24.
- If X_1 is near to 0.28 and X_2 is near to 0.046 and S is near to 1. 66 then pIC_{50} is near to 6.92.
- If X_1 is near to 0.26 and X_2 is near to 0.042 and S is near to 1. 66 then pIC_{50} is near to 4.66.

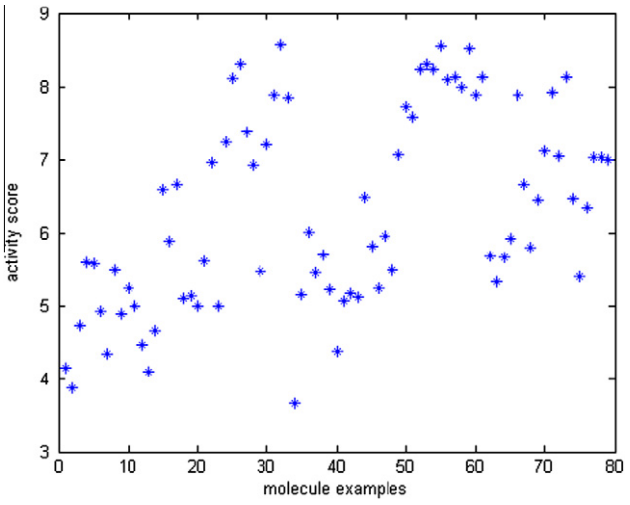

Fig. 4. Real values for pIC_{50} activity.

Table 2
Comparing the results of DT-ANFIS method with two other methods.

Method	Correlation coefficient	Number of selected variables
3D QSAR-PLS	0.931	6
ANN	0.97	4
Stepwise MLR-ANN	0.92	5
DT-ANFIS	0.935	3

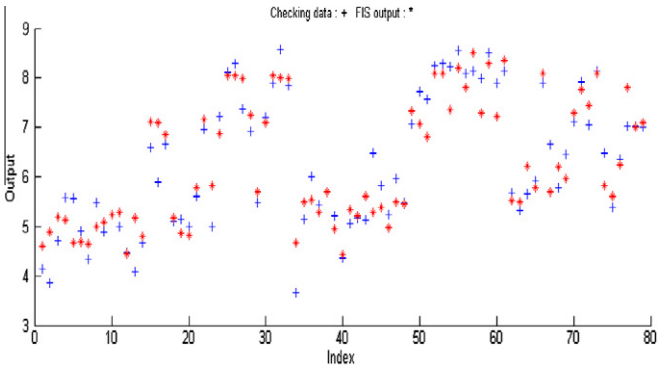

Fig. 5. Comparison of predicted values with real values for pIC_{50} activity.

- If X_1 is near to 0.25 and X_2 is near to 0.046 and S is near to 1. 66 then pIC_{50} is near to 5.17.
- If X_1 is near to 0.29 and X_2 is near to 0.074 and S is near to 1.65 then pIC_{50} is near to 7.11.
- If X_1 is near to 0.29 and X_2 is near to 0.148 and S is near to 1.71 then pIC_{50} is near to 6.35.

The predict activity (pIC_{50}) by ANFIS and real activity for each examples are compared in Fig. 5. In this figure, the real activity for checking data and the predict activity training data are same degrees with minor error. It proved that the DT-ANFIS method used in this paper is feasible and could be used to predict the pIC_{50} activity. The compared points seem to be close to each other indicating good agreement.

4.4. Comparison of hybrid DT-ANFIS model with stepwise MLR, ANN and 3D QSAR-PLS

For more investigation, stepwise MLR-ANN (Bazoui et al., 2002), 3D QSAR-PLS (Ravichandran et al., 2009) and ANN (Douali et al., 2004; Jalali-Heravi & Parastar, 2000) techniques are used to select the most important descriptors and predicting the anti-VIH activity. To find the best model, MLR and PLS were run sevral times with different settings of initial populations. The best models with stepwise MLR and PLS with best fitness were selected. The selected descriptors appeared in these models were used in developing NN to predict the values of pIC_{50}. The results are summarized in Table 2. It is clear from this table that the results of DT-ANFIS are superior compared with those of stepwise MLR-NN, 3Q QSAR-PLS and NN. The coefficient correlation is high and DT-ANFIS model uses fewer descriptors for predicting the inhibition activities.

5. Conclusion

The aim of the present work was developing a QSAR model to predict the inhibitory activities of anti-VIH molecules. However, a very important step in every QSAR studies is selecting suitable descriptors using a variable selection method. In this work, the DT algorithm technique has been employed for variable selection and satisfactory results have been obtained. Selection of three variables of X_1, X_2 and S by DT algorithm with electronic, constitutional, geometric and empirical characteristics indicates the complexity of inhibition mechanism. It is shown in this work that the DT as a variable selection method combined with ANFIS as a mapping tool can successfully solve this problem. Results show the satisfactory performance of DT–ANFIS models proposed in reducing the prediction error of the anti-VIH activity. Comparing the results of stepwise MLR, ANN and 3D QSAR methods with those

ofr DT-ANFIS reveals that the latter model selects the best variables to predict the inhibition action of anti-VIH molecules.

References

Bazoui, H., Zahouily, M., Boulaajaj, S., Sebti, S., & Zakarya, D. (2002). QSAR for anti-HIV activity of HEPT derivatives. *SAR and QSAR Environmental Research, 13*, 567–577.

Carne, A. C. (1987). ABC of AIDS, neurological manifestations. *British Medical Journal, 294*, 1399–1401.

Chiu, S. L. (1994). Fuzzy model identification based on cluster estimation. *Journal of Intelligent & Fuzzy Systems, 2*, 267–278.

Cho, S. J., & Hermsmeier, M. A. (2002). Genetic algorithm guided selection: variable selection and subset selection. *Journal of Chemical Information and Computer Sciences, 42*, 927–936.

Douali, L., Cherqaoui, D., & Villemin, D. (2004). Exploring QSAR of non-nucleoside reverse transcriptase inhibitors by neural networks: TIBO derivatives. *International Journal of Molecular Science, 5*, 48–55.

Garg, R., Gupta, S. P., Gao, H., Babu, M. S., & Debnath, A. K. (1999). Comparative quantitative structure–activity relationships studies on anti-HIV drugs. *Chemical Reviews, 99*, 3525–3601.

Ghafourian, T., & Cronin, T. D. M. (2005). The impact of variable selection on the modelling of oestrogenicity. *SAR and QSAR in Environmental Research, 16*, 171–190.

Jalali-Heravi, M., & Parastar, F. (2000). Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *Journal of Chemical Information and Computer Sciences, 40*(1), 147–154.

Jang, J. S. R. (1993). ANFIS: Adaptive-network-based fuzzy inference systems. *IEEE Transactions on Systems Man and Cybernetics, 23*(3), 665–685.

Kissi, M., Ramdani, M., Bouchon-Meunier, B., & Zakarya, D. (2008). Pattern recognition system based on empirical knowledge: Sandalwood and camphoraceous odours application. *Mathematics and Computers in Simulation, 77*, 453–463.

Kissi, M., Ramdani, M., Tollabi, M., & Zakarya, D. (2004). Determination of fuzzy logic membership functions using genetic algorithms: Application to structure-odor modelling. *Journal of Molecular Modeling, 10*, 335–341.

Marsala, C. (1998). *Apprentissage inductif en présence de données imprécises: Construction et utilisation d'arbres de décision flous*. Thèse de Doctorat, Université Pierre et Marie Curie, Paris, France.

Marsala, C., Ramdani, M., Tollabi, M., & Zakarya, D. (2000). Fuzzy decision trees to extract features of odorous molecules. In B. Bouchon-Meunier, R.R. Yager, L.A. Zadeh (Eds.), *Uncertainty in intelligent and information systems, advances in fuzzy systems: Applications and theory series* (pp. 235–249).

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*(1), 86–106.

Ramdani, M. (1994). *Système d'induction formelle à base de connaissances imprécises*. Thèse de Doctorat, Université Pierre et Marie Curie, Paris, France.

Ramdani, M., Kissi, M., & Bouchon-Meunier, B. (2004). Man–machine interaction to extract features of odorous molecules. In D. Ruan & X. Zeng (Eds.), *Intelligent sensory evaluation: Methodologies and applications* (pp. 255–268).

Ravichandran, V., Prashantha Kumar, B. R., Sankar, S., & Agrawal, R. K. (2009). Predicting anti-HIV activity of 1,3,4-thiazolidinone derivatives: 3D-QSAR approach. *European Journal of Medicinal Chemistry, 44*(3), 1180–1187.

Redfield, R., Blattner, W., & Gallo, A. R. C. (1999). In T. C. Merigan, Jr., J. G. Bartlett, & D. Bolognesi (Eds.), *Textbook of AIDS medicine*. Williams & Wilkins.

Ros, F., Pintore, M., & Chretien, J. R. (2002). Molecular descriptor selection combining genetic algorithms and fuzzy logic: Application to data base mining procedure. *Chemometrics and Intelligent Laboratory Systems, 63*, 15–26.

Smith, M., & Spittle, M. (1987). ABC of AIDS, tumours. *British Medical Journal, 294*, 1274–1277.

Takagi, T., & Sugeno, M. (1985). Fuzzy identification of systems and its application to modelling and control. *IEEE Transactions on Systems, Man and Cybernetics, 15*, 116–132.

Tanaka, H., Takashima, H., Ubasawa, M., Sekiya, K., Nitta, I., Baba, M., et al. (1992). Structure–activity relationships of 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) analogues: Effect of substitutions at the C-6 phenyl ring and the C-5 position on anti-HIV-1 activity. *Journal of Medicinal Chemistry, 35*, 337–345.

Yasri, A., & Hartsough, D. (2001). Toward an optimal procedure for variable selection and QSAR model building. *Journal of Chemical Information and Computer Sciences, 41*(5), 1218–1227.

Zadeh, L. A. (1977). Fuzzy sets and their applications to classification and clustering. In J. Van Ryzin (Ed.), *Classification and clustering* (pp. 251–299). New York: Academic Press.