Classification and Regression Trees—Studies of HIV Reverse Transcriptase Inhibitors

M. Daszykowski,[†,‡] B. Walczak,[†,‡] Q.-S. Xu,[†] F. Daeyaert,[§] M. R. de Jonge,[§] J. Heeres,[§]
L. M. H. Koymans,[§] P. J. Lewi,[§] H. M. Vinkers,[§] P. A. Janssen,[§] and D. L. Massart*[,†]

FABI, ChemoAC, Vrije Universiteit Brussels, Laarbeeklaan 103, B-1090 Brussels, Belgium, and
Center for Molecular Design, Janssen Pharmaceutica N.V., Antwerpsesteenweg 37, B-2350, Vosselaar, Belgium

Received August 5, 2003

In this paper, the application of Classification And Regression Trees (CART) is presented for the analysis of biological activity of Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs). The data consist of the biological activities, expressed as pIC_{50}, of 208 NNRTIs against wild-type HIV virus (HIV-1) and four mutant strains (181C, 103N, 100I, 188L) and the computed interaction energies with the Reverse Transcriptase (RT) binding pocket. CART explains the observed biological activity of NNRTIs in terms of interactions with individual amino acids in the RT binding pocket, i.e., the original data variables.

INTRODUCTION

The increasing number of infections and the emergence of new mutants of the human immunodeficiency virus (HIV), responsible for the acquired immunodeficiency syndrome (AIDS), require efficient screening procedures, able to preselect potential Reverse Transcriptase inhibitors from a vast collection of available compounds. The knowledge about the HIV structure and its life cycle has led to the identification of several potential drug targets. One of these targets is HIV Reverse Transcriptase, RT. It is an asymmetric heterodimer, consisting of two subunits of different size. The RT transcribes the viral single-stranded RNA to double-stranded DNA, which only in this form can be incorporated into the genome of the infected cell. Once the viral information is introduced into the cell's genome, the cell becomes a donor of the key substrates necessary for HIV replication. Therefore, by inhibiting the biological role of RT, virus replication can be stopped. The biologically active molecules inhibit the RT by strong interactions with certain key regions (amino acids) within the RT. There are two groups of RT inhibitors, known as Nucleoside Reverse Transcriptase Inhibitors (NRTIs) and Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs), which bind to RT in different ways.

The studied data set contains 208 NNRTIs, for which the pIC_{50} against the wild-type of the HIV virus and four mutants were obtained by Tibotec-Virco (Mechelen, Belgium[1]). They bind to a hydrophobic pocket, located in the larger subunit of RT. Knowledge about the strength of possible interactions is essential for understanding the activity of a molecule. The interaction energies can be computed by docking candidate ligands (here NNRTIs) into the binding site of HIV-RT, using molecular mechanics.[2] These interactions consist of the Coulomb and van der Waals interactions of the ligand with the side-chain and backbone residues of the amino acids (residues) located in RT. Since different types of compounds can interact with the hydrophobic pocket with different strengths, it is important to understand which residues play a role in these interactions. Moreover, the occurrence of mutations, which allows the virus to develop resistance for a certain inhibitor, is related to particular residues in the RT pocket. By substituting them by other amino acids, the interactions with previously active inhibitors become weaker.[3] A data mining method, CART, was applied to the NNRTI data set. This method was chosen because it selects the variables that are the most meaningful to describe differences in the data. The method has until now been used only rarely in the context of drug discovery, and the main aim of this article is to present the method and illustrate it with a case study, i.e., the data set described higher. An overview of the applications of CART to different problems of data analysis can be found in refs 4−9.

THEORY

Classification And Regression Trees (CART). The Classification And Regression Trees, CART, is a nonparametric method for classification and regression.[4,10−12] Its aim is to find mutually exclusive regions of the data space containing homogeneous subsets of the data. The dependent variable can be either numerical or categorical, leading to regression or classification trees, respectively. The CART results are presented in the form of a binary decision tree, containing nodes connected by branches. Nodes giving a rise to two new nodes (child nodes) are called parent nodes, otherwise they are terminal nodes. The tree is constructed via a recursive procedure partitioning objects from a parent node into two child nodes. Each node is characterized by a logic rule, usually defined for a single exploratory variable, leading to two more homogeneous child nodes compared to

* Corresponding author phone: +32-2-477-4737; fax: +32-2-477-4735; e-mail: fabi@vub.vub.ac.be.
[†] Vrije Universiteit Brussels.
[‡] On leave from Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland.
[§] Janssen Pharmaceutica N.V.

10.1021/ci034170h CCC: $27.50 © 2004 American Chemical Society
Published on Web 01/13/2004

the parent node. The partitioning procedure is stopped when the similarity of objects within nodes is the highest possible or the nodes contain a minimal number of objects predefined by a user. The CART procedure consists of three main steps. First, a complete tree is grown by recursive data partitioning. Then, a set of smaller, nested trees is found by reducing the number of nodes in the tree (pruning). In the last step, the optimal complexity of the tree is selected from this set, taking into account its predictive abilities for new samples.

Growing Complete Trees. The CART procedure searches through all possible splits of explanatory variables to find the best candidate, dividing objects into two more homogeneous groups (nodes). The best split and its variable are called primary split and primary variable, respectively. The objects characterized by lower and higher values than a split value found for a primary variable are placed into left, t_L, and right, t_R, descendant nodes of the branch, respectively. The goodness of the split on a primary variable, v, is judged by a reduction of impurity function, ΔI, for the two child nodes of node t

$$\Delta I(v, t) = I(t) - p_L I(t_L) - p_R I(t_R) \qquad (1)$$

where $I(t)$ is the impurity of the node t, v is the explanatory variable being a potential candidate for a split at parent node t, and p_L and p_R correspond to the proportion of objects in node t, placed in left and right child nodes, respectively.

There are several possible impurity functions.[4] For regression trees, the most popular impurity function is the sum of squared deviations of the dependent variable from the mean or median of dependent variable of objects in the node. The impurity of a parent node t is defined as

$$I(t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2 \qquad (2)$$

Thus, the impurity reduction for a parent node t and its two child nodes t_L and t_R, according to eq 1, is given as

$$\Delta I(v, t) = \sum_{x_n \in t} (y_n - \bar{y}(t))^2 - p_L\left(\sum_{x_n \in t_L} (y_n - \bar{y}(t_L))^2\right) - p_R\left(\sum_{x_n \in t_R} (y_n - \bar{y}(t_R))^2\right)$$

where y_n denotes the value of the dependent variable for objects x_n contained in node t and $\bar{y}(t)$ is the mean or median of the dependent variable of objects x_n.

For classification trees, i.e., when a categorical dependent variable is used, the most popular impurity functions are the entropy, Gini and Twoing indices.[4] The entropy impurity index of node t, $I(t)$, is expressed as

$$I(t) = -\sum_{i=1}^{k} p_i(t) \ln(p_i(t)) \qquad (3)$$

where $p_i(t)$ is the proportion of objects belonging to the ith class in node t, and k corresponds to the number of classes in the data.

The entropy index describes the within-node variance. The Gini diversity index is designed from the standpoint that a pure node, i.e., containing only one class of objects, has no diversity. Then the Gini impurity index reaches its minimum. If a node contains the same number of objects from each class, then the Gini index has the highest value. This condition can be presented by the following equation

$$I(t) = 1 - \sum_{i=1}^{k} (p_i(t))^2 \qquad (4)$$

where $I(t)$ denotes the Gini impurity index of node t, and $p_i(t)$ is the proportion of objects from the ith class in node t.

The Twoing impurity index can be applied when there are more than two classes of objects in the studied data. The main idea behind this impurity index is that it tends to group similar classes of objects near the top of the tree and to isolate the individual classes in terminal nodes. The Twoing impurity index for node t, $I(t)$, is defined as

$$I(t) = \frac{p_L p_R}{4} \sum_{i=1}^{k} (|p_i(t_L) - p_i(t_R)|^2) \qquad (5)$$

where p_L and p_R denotes the proportion of objects in the left and right descendant nodes, respectively. The factor $1/4(p_L p_R)$ in eq 5 is introduced to favor the splits containing approximately the same number of objects in descendant nodes.

Using one of the impurity measures, the splitting procedure is continued until the impurity function reaches its minimum or the number of objects in the terminal nodes is smaller than a predefined value. In this way the complete tree is built.

Pruning Procedure. Although the complete tree fits the data well, it can have poor predictive abilities for new samples. To find the tree with optimal size, the tree branches are cut, making the tree smaller, as long as the accuracy of the smaller trees compared with larger ones remains satisfactory. This step of the method is called pruning. In our study the so-called cost-complexity pruning, proposed by Breiman et al.,[4] has been applied. In this approach the accuracy of the tree is presented as a sum of resubstitution error, $R(T)$, and a penalty component proportional to the tree size, $|\tilde{T}|$

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \qquad (6)$$

where $R_\alpha(T)$ is the cost-complexity measure, $R(T)$ is the within-node sum of squares (for regression trees) or misclassification error (for classification trees), $|\tilde{T}|$ is the number of terminal nodes in the subtree (T), and α is a complexity parameter.

The cost-complexity pruning relies on removing the weakest branches of the tree, starting from the terminal nodes, gradually increasing the complexity parameter α. If the complexity parameter equals 0, then the largest tree is considered. To find the weakest branch of the tree with a certain size, a smaller subtree is found with comparable resubstitution error to the larger tree. As proven by Breiman,[4] for a certain α value, among all subtrees of the same size, there exists only one subtree with minimal value of the cost-complexity measure. Thus, the α parameter is a measure of additional accuracy introduced by a new node.[4] In such a way a sequence of nested subtrees is obtained.

Selection of the Optimal Tree Size. The selection of the optimal tree complexity is the last step of the CART procedure. From a sequence of nested trees the optimal one is selected as the one offering the best predictive ability for new samples. In practice, to judge the predictive abilities of a tree an independent data set is required. If the original data set is large enough it can be divided into a model and a test set, which are used to construct and validate the tree, respectively. Alternative to this approach is V-fold cross-validation, which provides the "honest" estimates of the prediction error for new samples.[4] In this procedure, the data set is split into V subsets, each of which is characterized by a similar distribution of objects with respect to the dependent variable. One out of V subsets is used to obtain the prediction error of the tree constructed for the remaining V-1 subsets. The whole procedure is repeated V times, and V different trees are produced. The final prediction error is given as the average of all prediction errors computed for trees of certain complexity. The resubstitution and cross-validation errors are plotted against the size of the tree. With increasing size of the tree the resubsitution error decreases and reaches its minimum for the largest tree, which perfectly fits the data but may not have good prediction ability due to over-fitting. The cross-validation error at the beginning decreases with the size of the tree but later on starts to rise, which indicates that the prediction of the tree is getting worse for new objects. The optimal tree is the simplest one among trees characterized by a cross-validation error within one standard error of the minimal value of the cross-validation error (1-SE rule).[4]

Although the cost-complexity pruning proposed by Breiman et al.[4] seems to be the most popular, there are other pruning approaches, for instance pessimistic pruning and pruning based on Minimum Description Length,[13] and many more.[14]

Primary, Competitive, and Surrogate Splits. For every primary split, competitive splits can be evaluated based on the impurity reduction. The impurity reduction for the best competitor split is smaller than or the same as for the primary split. The first competitor would be taken as the primary split if the original primary split would be removed. However, the new tree constructed with the first competitor, will not necessarily have the same structure as the tree obtained by a primary split.

The surrogate splits aim to mimic as much as possible the result of the primary split. The ability of the surrogate split to reproduce the same split as a primary split is evaluated based on assignment of individual objects from a parent node to child nodes by primary and surrogate splits. Thus, if the content of child nodes of primary and surrogate splits are the same, the surrogate mimics perfectly the action of a primary split.

To explain how the surrogate splits are found, some definitions must be introduced. For any data object, the prior probability that it will be assigned to node t, $p(t)$, is given as

$$p(t) = \sum_{j=1}^{k} \pi_j \frac{N_j(t)}{N_j} \tag{7}$$

where π_j is the ratio between the number of objects in the jth class and the total amount of objects, $N_j(t)$ is the number of objects from the jth class in node t, and N_j denotes the number of objects from the jth class.

Objects split at node t are assigned to left or right descendant nodes (t_L, t_R), with the relative probabilities of assignment to the left, $p(t_L)$, or right descendant nodes, $p(t_R)$, given as

$$p(t_L) = \frac{1}{p(t)} \sum_{j=1}^{k} \pi_j \frac{N_j(t_L)}{N_j} \tag{8}$$

$$p(t_R) = \frac{1}{p(t)} \sum_{j=1}^{k} \pi_j \frac{N_j(t_R)}{N_j} \tag{9}$$

Both probabilities $p(t_L)$ and $p(t_R)$ are normalized by $1/p(t)$, to make their sum equal to 1. The split on node t can be viewed as a probability rule deciding to which child node an object should be assigned. This probability rule is given as the maximal value of the probability that an object will be placed in left or right child nodes. Suppose that the split t sends the objects to the left child node with the relative probability $p(t_L) = 0.7$ and to the right child node with probability $p(t_R) = 0.3$. Then, the probability rule of assignment is given as $\max(p(t_L),p(t_R))$, i.e., 0.7, whereas the error of this probability rule is equal to $1 - \max(p(t_L),p(t_R))$, i.e., 0.3.

Since the surrogate split, s', should mimic the action on the primary split, s, the probability that the surrogate split places the same objects in the left and right descendant nodes as the primary split is presented as the probability of agreement

$$p(t_L \cap t_{L'}) = \frac{1}{p(t)} \sum_{j=1}^{k} \pi_j \frac{N_j(\mathrm{LL})}{N_j} \tag{10}$$

$$p(t_R \cap t_{R'}) = \frac{1}{p(t)} \sum_{j=1}^{k} \pi_j \frac{N_j(\mathrm{RR})}{N_j} \tag{11}$$

where $p(t_L \cap t_{L'})$ and $p(t_R \cap t_{R'})$ are the probabilities of agreement that the primary and surrogate splits place the same objects in the right and left nodes, respectively, and $N_j(\mathrm{LL})$ and $N_j(\mathrm{RR})$ are the number of the same objects sent to the same nodes (left or right) by primary and surrogate splits.

The probability that the surrogate split behaves as the primary split is given by the sum of the agreement probabilities

$$p(s,s') = p(t_L \cap t_{L'}) + p(t_R \cap t_{R'}) \tag{12}$$

where $p(s,s')$ denotes the agreement probability, describing to what extent the surrogate split mimics the action of the primary split.

For a given primary split the agreement probabilities between the primary split and its potential surrogate splits are calculated. Then, the probability agreement is compared to the error of the probability assignment rule computed for the primary split. This measure is called association between a surrogate and the primary split, $\lambda(s,s')$. It is defined as

$$\lambda(s,s') = \frac{\min(p_L(t),p_R(t)) - (1 - p(s,s'))}{\min(p_L(t),p_R(t))} = 1 - \frac{(1 - p(s,s'))}{\min(p_L(t),p_R(t))} \tag{13}$$

where $\min(p_L(t),p_R(t))$ is the error of the assignment rule observed for the primary split s on node t, and $p(s,s')$ is given by eq 12.

The association equals 1 if the potential surrogate split reproduces exactly the same two nodes, object by object, as the primary split. Otherwise, the association is smaller than 1. Only the splits on variables with an association higher than zero are surrogate splits. In practice, if the assignment rule for a primary split is very strong, i.e., the error of the assignment rule is small, the potential surrogate should show very high agreement with the primary split in order to obtain high association. It may happen that the association measure is small, although the surrogate split agrees with the primary split on 95% of cases. For instance, if the error of the assignment rule for the primary split equals 0.1, and the probability of agreement between the primary and surrogate splits equals 0.95, then the association equals 0.5. For the same primary split, if the agreement probability between the primary and surrogate splits equals 0.98, then the association equals 0.8.

The concept of the surrogate splits plays an important role in CART method. The surrogate splits can help to deal with missing elements, by substituting missing elements with the best surrogate split. Moreover, the surrogate splits are used to evaluate the importance of the variables.

Importance of Variables. The importance of a variable is evident if it occurs in the tree at one of its branches. However, considering the primary split variables as the only important ones is misleading, because other variables may give splits with comparable quality. After removing the primary variable, the first available competitor variable becomes a new primary variable, which can lead to a tree with similar accuracy as the previous one. For this reason, the primary splits variables act as masking variables. To score the importance of the variable v in modeling the dependent variable, the CART procedure searches the surrogate split with the highest association. The importance of the variable v is given as an impurity reduction caused by replacing the primary split by its best surrogate split. The importance of variables is scored between 0 and 100. It gives information about the degree of masking in the tree and the potential splitting power of the other variables. However, this measure is strictly related to a particular tree. Removing a variable from the tree causes other splits, with a new list of surrogate splits. Since the variable importance is scored taking into account possible surrogates, other surrogates yield a different scoring of variables. Therefore, the variable importance is specific for one particular tree.

DATA

The data set consists of the inhibitory activities of 208 NNRTIs and their calculated van der Waals and Coulomb interactions with the side-chain and backbone parts of a selected set of 93 amino acid residues lining the NNRTI binding pocket. The nonbond interaction energies between the NNRTIs and the RT binding pocket were calculated using a pharmacophore based docking algorithm.[15] Predicting the structure of an inhibitor-binding site complex is a nontrivial problem. Moreover, we have found evidence for multiple binding modes of NNRTIs.[16] We therefore have used the following procedure to calculate the nonbond interaction energies. First, a set of 563 NNRTIs was selected of which the inhibitory activities against wild-type HIV and the mutant strains 181C, 103N, 100I, and 188L have been determined. These NNRTIs were docked into a panel of 16 binding pockets, obtained from the crystal structures of HIV-RT complexed with several inhibitors. The crystal structures are listed in Table 1.

Table 1. Crystal Structures of Reverse Transcriptase − An Overview

no.	name	origin	ref
1	atu	proprietary	
2	1bqm	Protein Data Base	20
3	1c1b	Protein Data Base	21
4	1c1c	Protein Data Base	21
5	dat	proprietary	
6	1dtq	Protein Data Base	22
7	1dtt	Protein Data Base	22
8	1fk9	Protein Data Base	23
9	1hnv	Protein Data Base	24
10	1rev	Protein Data Base	25
11	1rt1	Protein Data Base	26
12	1rt2	Protein Data Base	26
13	1rt4	Protein Data Base	27
14	1rt5	Protein Data Base	27
15	1rt6	Protein Data Base	27
16	1rti	Protein Data Base	28

The pharmacophore docking algorithm[15] is based on the presence of a common pharmacophore in the NNRTIs studied. The pharmcophore consists of a hydrophobic center that interacts with a hydrophobic pocket formed by tyrosine A181, tyrosine A188, and tryptophane A229 and a hydrogen bond acceptor and a hydrogen bond donor that interact with the lysine A101 backbone. The ligands are positioned into the pocket by mapping all combinations of pharmacophore points of all low-energy conformations of a ligand onto the pharmacophore of the native ligand in the crystal structure, as illustrated in Figure 1.

After a first energy evaluation of the interaction energy between the putative ligand and the pocket, a full molecular mechanics energy minimization is carried out on the pharmacophore/conformer combinations with the lowest energy. The minimized ligand−pocket complex with the lowest nonbond energy between the ligand and the pocket is retained as the final solution of the docking.

The nonbond interaction energy is calculated as the sum of the van der Waals and Coulomb interactions between the ligand and the pocket atoms. The calculated nonbond interaction energies are not to be confounded with the free energies of binding. Calculating the latter would involve the assessment of entropic contributions and solvation issues, requiring computationally far more involved methods such as molecular dynamics or Monte Carlo simulations.[2] The nonbond interaction energies should therefore not be seen in a thermodynamic context but as molecular descriptors.

In the MMFF94 force field, the van der Waals interaction is given by[17]

$$E_{\mathrm{vdW}} = \sum_{i}^{N_{\mathrm{pocket}}} \sum_{j}^{N_{\mathrm{ligand}}} \epsilon_{ij} \left(\frac{1.07 R^{*}_{\mathrm{IJ}}}{r_{ij} + 0.07 R^{*}_{\mathrm{IJ}}} \right)^{7} \left(\frac{1.12 R^{*\,7}_{\mathrm{IJ}}}{r_{ij}^{\;7} + 0.12 R^{*\,7}_{\mathrm{IJ}}} - 2 \right)$$

where r_{ij} is the Euclidean distance between pocket atom i and ligand atom j, R^{*}_{IJ} is the minimal-energy separation between the atom types I and J, and ϵ_{ij} is the well-depth.

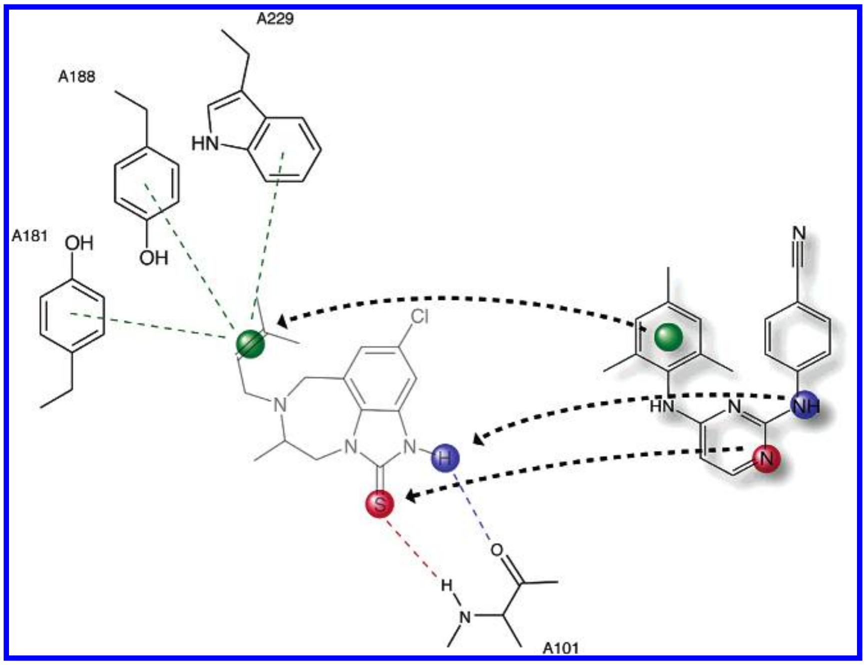

Figure 1. Positioning of a ligand into a binding pocket by mapping of the pharmacophore points.

The Coulomb interaction is given by

$$E_{\text{Coulomb}} = \sum_{i}^{N_{\text{pocket}}} \sum_{j}^{N_{\text{ligand}}} \frac{332.0716 q_i q_j}{D(r_{ij} + 0.005)^n}$$

where q_i and q_j are the atomic partial charges, D is the dielectric constant, and the exponent n allows the dielectric constant to be distance-dependent. We used a distance-dependent dielectric constant ($n = 2$) and a value of 4 for D.

Of the 563 NNRTIs, 208 where found to bind in at least 10 of the 16 binding pocket structures with a nonbond interaction energy smaller than −30 kcal/mol. Of these 208 NNRTIs, the van der Waals and Coulomb interactions with the side-chain and backbone parts of the 93 amino acids forming the binding pockets were averaged, leading to a 208 × 372 (2 × 2 × 93) table. Of this table, the 54 columns with an average absolute contribution of more than 0.1 kcal/mol were selected. The final data set consists of this 208 × 54 matrix of calculated energy contributions and the vector of the average of the biological activities, expressed as pIC_{50}, against wild-type HIV and the mutant strains 181C, 103N, 100I, and 188L. The pIC_{50} values were obtained by Tibotec-Virco (Mechelen, Belgium[18]). The 208 NNRTIs belong to five different chemical classes of compounds, denoted as TIBO (2 structures), HEPT (92 structures), ITU (1 structure), DATA (65 structures), and DAPY (48 structures).

The data set, containing the pIC_{50} values against the wild type and mutant HIV strains and the averaged calculated interaction energies, is provided as Supporting Information in a text format.

RESULTS AND DISCUSSION

Regression Tree. To trace which residues play an important role in describing biological activity of the NNRTIs, a regression tree was built. The splits are guided by the average activity of the NNRTIs (dependent variable) against the wild type of the virus and four mutants. As the impurity function the sum of squared deviations of the dependent variable from the mean of dependent variable of objects in the node was used (see eq 2). After cross-validation an optimal tree, containing 11 terminal nodes, was selected. The resubstitution error of the regression tree equals 0.32, whereas the cross-validation error is about 0.78. This indicates that for regression purposes the use of the regression trees is rather limited. Therefore, the regression tree is used here in an exploratory context only, i.e., to understand the importance of the interactions between the NNRTIs and the binding site of the RT pocket. The NNRTIs with pIC_{50} higher than 7.19 are denoted as highly active, and those with pIC_{50} smaller than 7.19 are denoted as less active (see Figure 2).

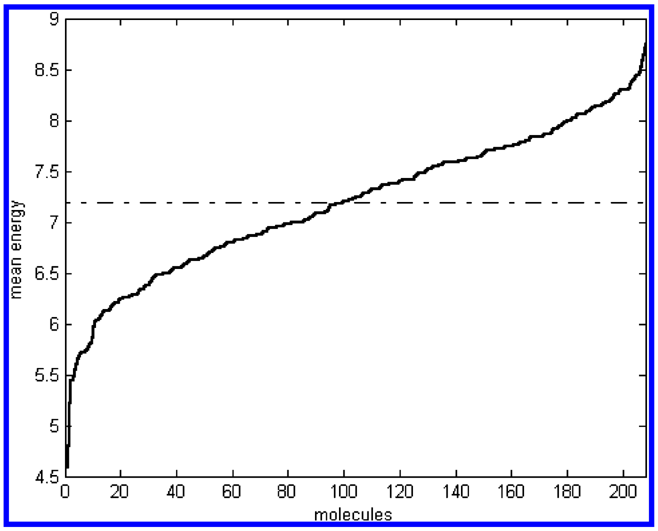

Figure 2. 208 NNRTIs sorted according to their biological activity. The dashed line is the mean value of the overall average biological activities (expressed as pIC_{50}) of NNRTIs against the wild-type and four mutant strains of the HIV virus.

In Figure 3a, the nodes containing mostly highly active compounds are marked with an 'H', and the nodes containing

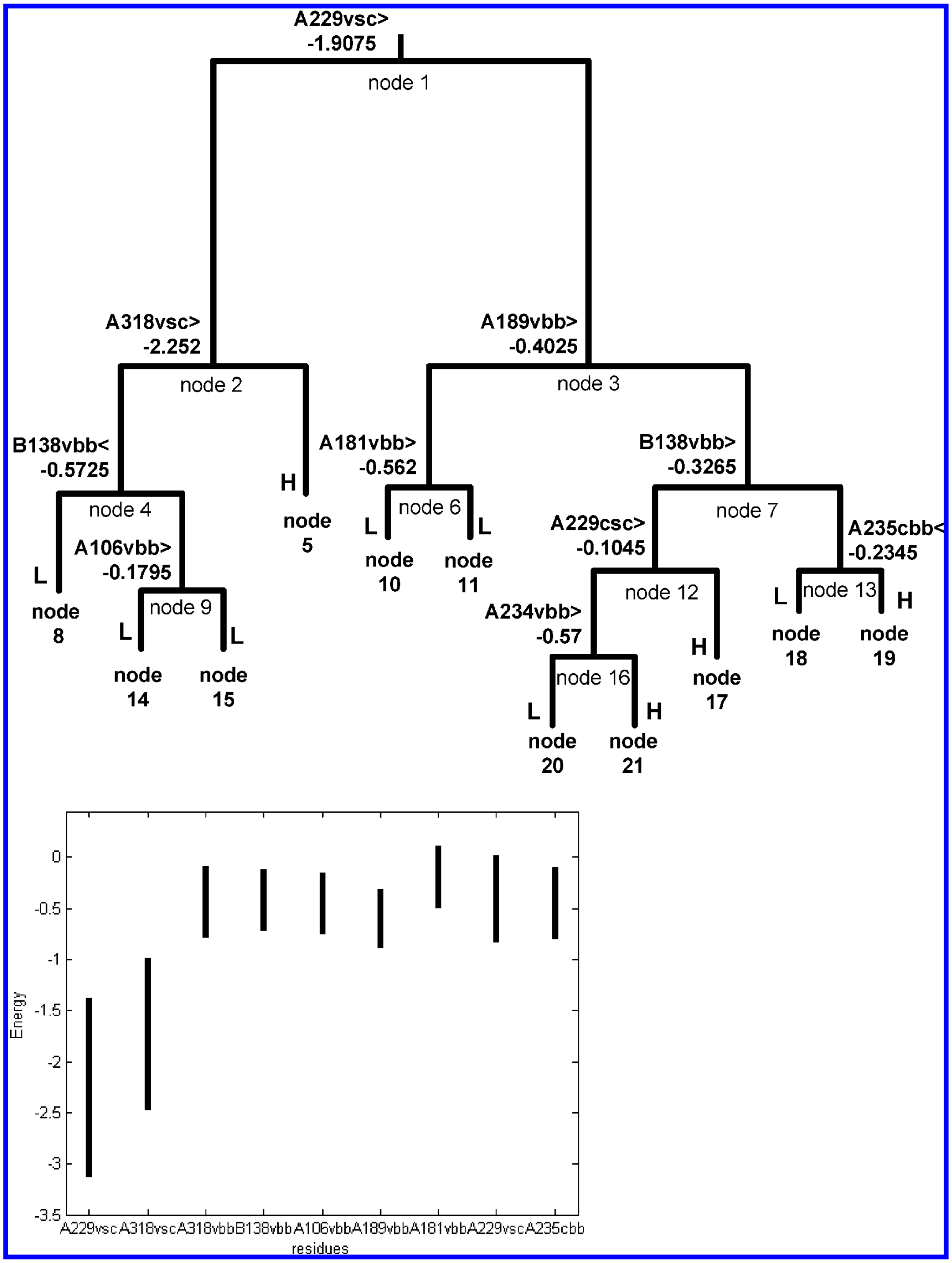

Figure 3. (a) Regression tree constructed for 208 NNRTIs with the target variable describing their biological activities. 'H' and 'L' denote nodes containing majority of highly or less active NNRTIs. (b) Energy range of certain interactions associated with primary splits selected by CART.

mostly less active compounds are marked with an 'L'. The regression tree constructed for 208 NNRTIs is presented in Figure 3a, whereas Figure 3b displays the energy range of the primary splits. A description of the dependent variable associated with the objects in each terminal node is given in Table 2.

The first split of the regression tree is by the van der Waals side-chain interaction of tryptophane A229 with the NNRTIs—see Figure 3a. This residue is situated at the 'roof' of the RT binding site and is highly conserved in HIV-RT.[24]

The activity of the highly active molecules is explained by their strong interaction with this residue (Figure 3b). Because of the important impurity reduction by this split, only a few consecutive primary splits are necessary to explain the relation between binding energies and the activity. The three consecutive competitors of this split are lysine A102, phenylalanine A227, and tyrosine A318 providing the Coulomb backbone, van der Waals backbone, and van der Waals side-chain interactions, respectively. The three first surrogate splits of the split on tryptophane A229 are phenylalanine A227 and leucine A234, both interacting through the backbone van der Waals interactions, and proline A95 interacting by the van der Waals interaction through the side chain. The surrogate splits have association measures

Table 2. Summary of the Terminal Nodes of the Regression Tree Constructed for 208 NNRTIs and Their Biological Activities[a]

terminal node	no. of objects	mean	activity	standard deviation	minimum	maximum
8	12	6.05	L	0.30	5.44	6.30
14	5	5.98	L	0.37	5.46	6.38
15	71	6.62	L	0.44	5.68	7.84
5	9	7.80	H	0.49	6.69	8.74
10	3	5.73	L	1.03	4.58	6.54
11	8	7.00	L	0.41	6.48	7.70
20	19	6.74	L	0.47	5.74	7.56
21	9	7.63	H	0.37	6.86	8.14
17	20	7.75	H	0.34	7.08	8.40
18	6	7.17	L	0.51	6.66	8.08
19	46	7.90	H	0.32	7.36	8.56

[a] The dependent variable of the molecules in the terminal nodes is described by its mean, standard deviation, and minimal and maximal values; ‘H’ and ‘L’ denote nodes containing a majority of highly active or less active compounds.

Table 3. Summary of the Terminal Nodes of the Classification Tree Constructed for 208 NNRTIs[a]

terminal node	no. of objects	less active	active
2	78	**60**	18
6	14	**14**	
7	8	2	**6**
10	13	**12**	2
11	4		**4**
9	91	12	**79**

[a] Bold characters denote terminal nodes with a majority of highly active or less active NNRTIs.

equal to 0.779, 0.716, and 0.712, respectively. Generally, the split on tryptophane A229 already partitions the molecules into active and less active molecules (see Figure 3a). The distribution of activity in node 2 shows that less active molecules are most prevalent (see Figures 2a and 3a), while in node 3 the distribution of the activity is skewed toward higher activity (see Figures 2a and 3b).

Although most of the molecules in node 2, i.e., with an interaction energy with the side chain of tryptophane A229 higher than −1.91 kcal/mol, are less active, a number of molecules have high activity. To distinguish them from the less active molecules, another split is necessary (see Figure 3). It is made on tyrosine A318. This residue is also highly conserved and resistant against mutation. Now, all misplaced active molecules belong to terminal node 5 and reveal strong van der Waals interactions, less than −2.25 kcal/mol, with the side chain of tyrosine A318. Three constitutive competitors for this split are glutamic acid B138, leucine A100, and lysine A102 interacting through the van der Waals backbone, van der Waals backbone, and Coulomb backbone, respectively. The surrogate splits found for tyrosine A318 are the van der Waals backbone, van der Waals side chain, and van der Waals backbone interactions with phenylalanine A227, proline A225, and lysine A101, respectively. The surrogate splits have high association measures equal to 0.949, 0.938, and 0.938, respectively.

The splits on tryptophane A229 and tyrosine A318 characterized by weak and strong interactions, respectively, gather 88 less active molecules in node 2. To explain the activity deviations in node 2, other splits are made on glutamine B138 and valine A106, respectively, leading to terminal nodes 8, 14, and 15 (see Figure 3a). In these nodes, some misclassifications can be noticed. For instance, terminal node 5 contains 8 out of 9 highly active molecules, which do not strongly interact with the side chain of the tryptophane A229 but interact with tyrosine A318 − see Figure 3. The standard deviation of the activity observed within node 5 is influenced by one less active molecule and equals 0.49 (see Table 2). In the terminal node 15 there are also a few highly active molecules (see Table 2). The misclassifications are caused by the synergetic effect of several components. The first is due to the lack of a clear border between highly active and less active molecules (see Figure 2). Several inhibitors have an activity close to the cutoff value. To make a better distinction between highly active and less active molecules, more splits are required. By growing a larger tree, more artifacts can be explained. However the interpretation of very large trees is difficult. Another source of misclassifications is the uncertainty of measured activities, being about 1 unit in log scale. Finally, the docking procedure can fail to generate a reasonable structure of the inhibitor−binding site complex. When validating the pharmacophore docking algorithm by cross docking a number of NNRTIs in each others binding pocket, it was found that 20% of the dockings produced a structure with a root-mean-square deviation of more than 3 Å from the experimentally determined structure.[15] Even after averaging the interaction energies over a number of dockings in several binding pockets, wrongly docked molecules can strongly change the overall picture of interactions and increase the misclassification rate.

The second branch of the tree is more complex than the first, and more splits are required to explain the presence of subgroups of molecules located in seven terminal nodes (see Figure 3a). The majority of highly active molecules are located in terminal node 19 (see Table 2). These molecules are characterized by strong interactions with RT through the van der Waals side-chain interaction of tryptophane A229, the van der Waals backbone interaction of valine A189, the van der Waals backbone interaction of glutamic acid B138, and the Coulomb backbone interaction of histidine A235− see Figure 3a. The other splits in the regression tree are less easily interpreted, since the corresponding interaction energies are much smaller (see Figure 3b).

Classification Trees. We have also used CART to construct classification trees of the NNRTIs. To construct the first classification tree, the 208 NNRTIs were divided into a group of highly active and a group of less active compounds. Hereby, the target variable guiding the splits consists only of zeros (less active NNRTIs) and ones (highly active NNRTIs). The same definitions of highly active and less active molecules are used as for constructing the regression tree. For the tree construction the Gini index was used. The classification tree is presented in Figure 5a, together with the energy ranges of interactions selected as the primary splits−Figure 5b. The detailed information about the content of the terminal nodes is given in Table 3.

The optimal classification tree is simpler than the regression tree and consists of six terminal nodes (see Figure 5a). Its resubstitution and prediction errors equal 0.25 and 0.67, respectively. The first primary split, as in the regression tree, is made on the tryptophane A229 residue−see Figure 5a.

The first split on tryptophane A229 gathers the majority of less active molecules (50 out of 91) in terminal node 2,

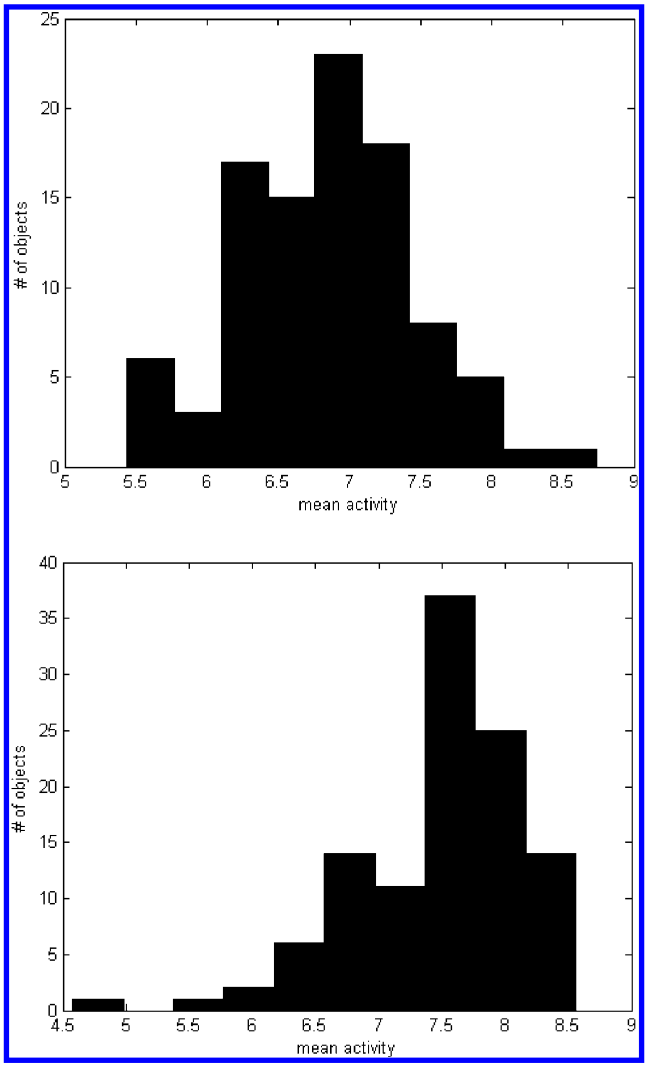

Figure 4. Histograms of the activity distribution in (a) left and (b) right main branch of the regression tree.

with 18 molecules wrongly classified (see Table 3). As was observed for the regression tree, the molecules in the terminal node are characterized by weaker interactions with the side chain of tryptophane A229, with the split occurring at −1.87 kcal/mol. It was found that some wrongly classified objects can form pure nodes if the tree is grown larger. However this does not entirely eliminate the misclassification problem. Upon visual inspection of the calculated complexes, it was found that two of the 18 wrongly classified molecules were docked in a highly improbable orientation in several pockets of RT. The remaining misclassification can be explained as for the regression tree, i.e., by the ‘hard’ division into highly active and less active molecules. The competitors of the tryptophane A229 split are the following: leucine A100, leucine A228, and proline A95. These residues form van der Waals interactions with the side chain, backbone, and side chain, respectively. The three surrogate splits of the tryptophane A229 split are phenylalanine A227, leucine A228, and tyrosine A318 providing van der Waals interactions with backbone, backbone, and side chain, respectively. Their association with the tryptophane A229 split is 0.803, 0.755, and 0.726, respectively.

The NNRTIs that strongly interact with the tryptophane residue A229, less than −1.87 kcal/mol, are placed in node 3, and most are highly active molecules (91 out of 109). The second split is made on the arginine residue A100, strongly interacting with the molecules by its van der Waals side chain. Contrary to the first split, the second split groups highly active compounds in node 5 at higher energies, i.e., more than −3.30 kcal/mol. Together with the residue in the first split, this residue is characterized by high interaction compared to other residues in the tree (see Figure 5b). The three competitors to the split on arginine A100 are tyrosine A188, histidine A235, and arginine A100. The surrogate splits are tyrosine A188, arginine A100, and proline A236, with association measures equal to 0.892, 0.885, and 0.877, respectively.

Another split occurs again on the tryptophane residue A299, this time at the van der Waals backbone interaction, but at a lower split value than those observed for the side-chain interaction. The terminal node 6 contains 14 less active NNRTIs. They are characterized by strong interactions with the side chain of tryptophane A229 and arginine A100 and weak interaction with tryptophane A229 backbone. The terminal node 7 is not pure and formed by two less and six highly active molecules. The split on tyrosine A188 groups in node 8 the majority of less active molecules belonging to node 3. Additionally these molecules strongly interact with the backbone of tyrosine by van der Waals interaction. The NNRTIs with a strong interaction with tryptophane A229 and tyrosine A188 are placed in terminal node 9. This node contains the majority of highly active compounds (see Table 3).

To understand the differences in interactions of NNRTIs belonging to different chemical classes, another classification tree was built, again with the Gini index as an impurity measure. This time, the splits are guided by the variable describing the chemical class of the compounds (TIBO, HEPT, ITU, DATA, or DAPY). The optimal classification tree contains six terminal nodes. The resubstitution and cross-validation errors are 0.09 and 0.26, respectively. The classification tree is shown in Figure 6a, and the energy range characteristic for the primary splits in Figure 6b. The information about the content of terminal nodes is summarized in Table 4. The first split on the arginine residue A100, split 1, separates the HEPT type compounds from the rest by placing them in terminal node 2 (see Figure 6a and Table 4).

The interaction between the HEPT type compounds is less strong than for the DATA or the DAPY compounds, due to the indirect interaction with the backbone of arginine A100 through a water molecule.[26] The DATA and DAPY compounds are able to create a direct hydrogen bond with the backbone part of arginine A100. Three main competitors of the arginine A100 split are as follows: lysine A101, proline A236, and arginine A100. Here, arginine A100 is at the same time a primary split and a competitor but at different threshold values. The three surrogate splits of the arginine A100 split are as follows: lysine A101, arginine A100, and proline A236, with association measures equal to 0.993, 0.947, and 0.947, respectively.

Another split distinguishes DATA and DAPY type compounds in node 3, by means this time of Coulomb interaction with the backbone part of arginine A100 (see Figure 6a and Table 4). The selection of this interaction can be explained by the differences in the chemical structure of the two classes of compounds. The DATA compounds contain an additional

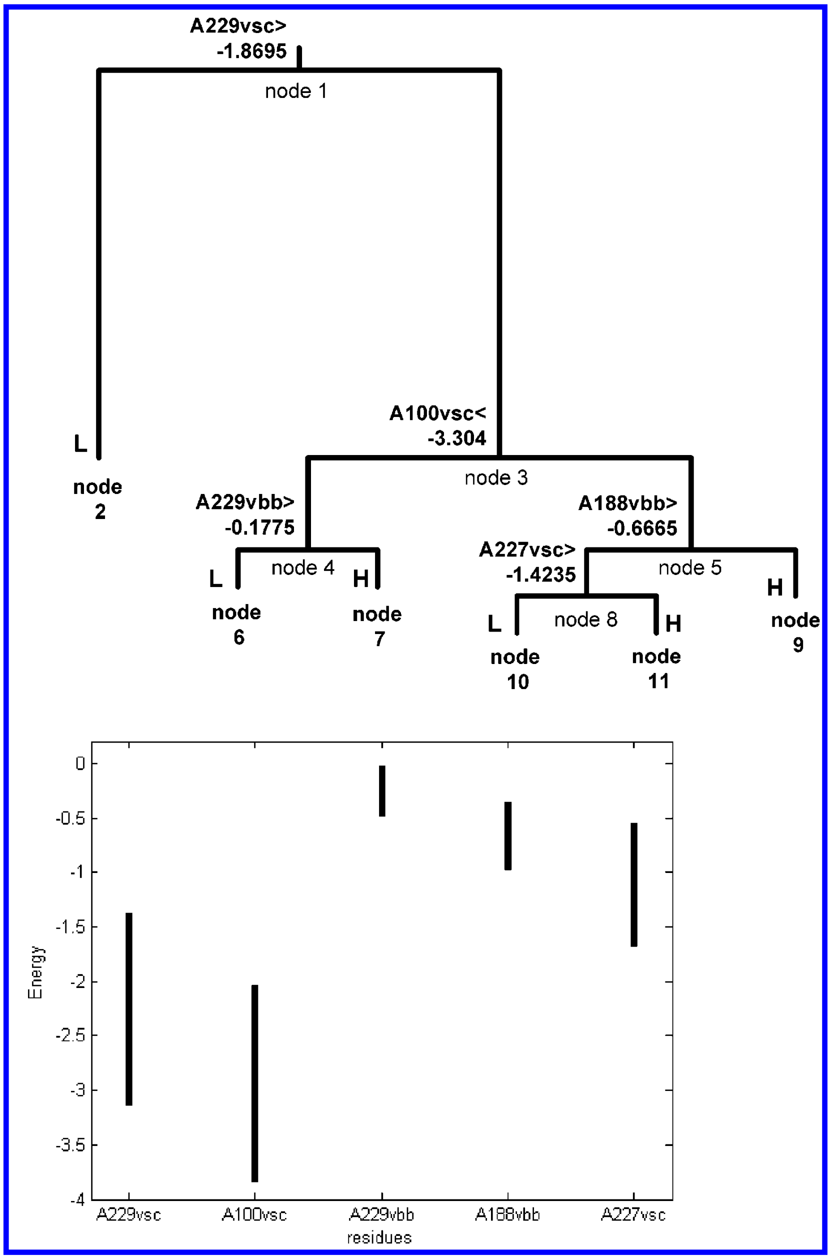

Figure 5. (a) Classification tree constructed for 208 NNRTIs with a binary target variable describing highly active ('H') and less active ('L') compounds and (b) energy range of selected interactions associated with primary splits selected by CART.

nitrogen atom with negative partial charge close to the arginine A100 backbone, which makes their Coulomb interaction with this residue weaker. The DATA compounds are placed in terminal node 6, whereas the DAPY compounds are placed in terminal node 9. The remaining classes of molecules, TIBO and ITU, are located in the terminal nodes 2 and 9.

The insights gained from the CART analysis are valuable in the design of novel NNRTIs using traditional medicinal chemistry approaches. Moreover, the automatic prediction of activities can be used as a scoring function for automatic in silico drug design algorithms (see e.g. ref 19).

CONCLUSIONS

The variables selected by CART are indeed important for the inhibition of RT. Most of these residues are located at the upper hemisphere and not at the entrance of the hydrophobic RT pocket (see Figure 7).

The majority of the interactions pointed out by CART are hydrophobic (van der Waals interactions), which is to be expected, since the NNRTI pocket is also hydrophobic. CART selects as the primary split the tryptophane A229 residue. Its biological importance in the inhibition process is illustrated by the fact that this residue is highly conserved.

It can be concluded that CART does allow for extracting information about the importance of variables for the inhibition of RT. Although the regression tree, due to its high prediction error, is not fit for prediction purposes, its information content gives an overall picture of the importance of the variables for explaining the observed biological activity in terms of the energies of different interactions. More generally, we conclude that CART is a useful tool for data mining in drug discovery. Depending on the type of information represented by a target variable, CART can give very simple answers in terms of the original data variables. Because we wanted to highlight the possibilities of CART,

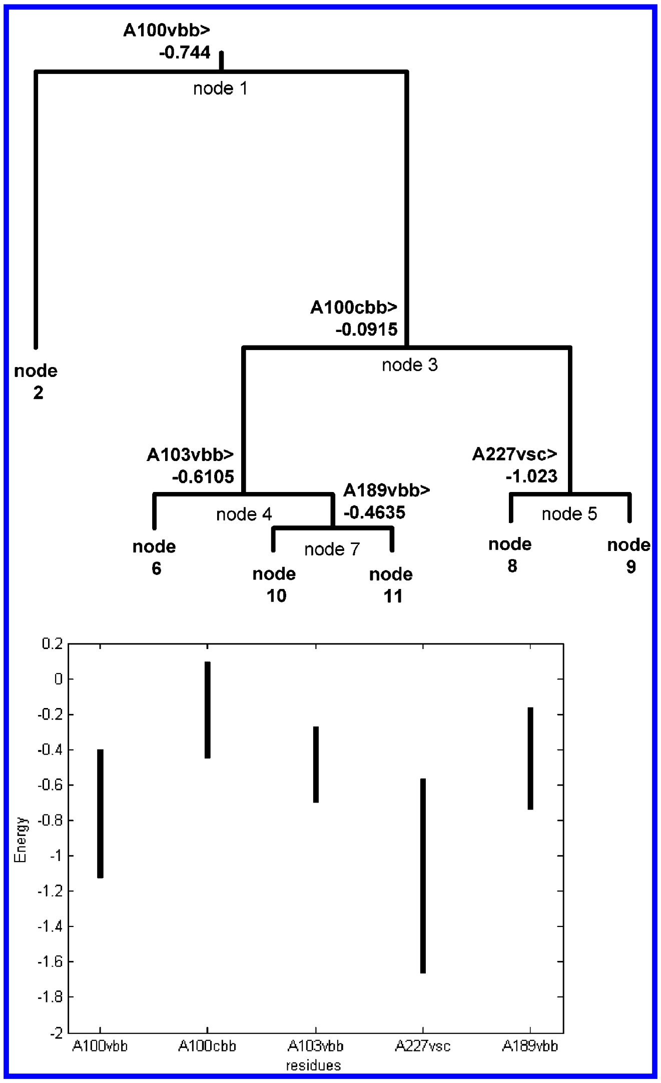

Figure 6. (a) Classification tree constructed for 208 NNRTIs with target variable describing the chemical class of the compounds (TIBO, HEPT, ITU, DATA, or DAPY) and (b) energy range of selected interactions associated with primary splits selected by CART.

Table 4. Summary of the Terminal Nodes of the Classification Tree Constructed for 208 NNRTIs Belonging to Five Chemical Classes of Compounds (TIBO, HEPT, ITU, DATA, and DAPY)[a]

terminal node	no. of objects	TIBO	HEPT	ITU	DATA	DAPY
2	*96*	2	**91**	1	2	
6	*48*				**48**	
10	*5*				5	
11	*5*					5
8	*8*				**7**	1
9	*46*		1		3	**42**

[a] Bold characters denote nodes containing a majority of NNRTIs belonging to one chemical class of compounds.

we did not present results of other exploratory data analysis techniques in this article. This may create the impression that we consider that CART should be used as the only technique in applications as the one described here. In fact, it is our opinion that different techniques such as Principal Component Analysis, clustering techniques, etc. should be used together with CART to understand as many aspects of the data as possible.

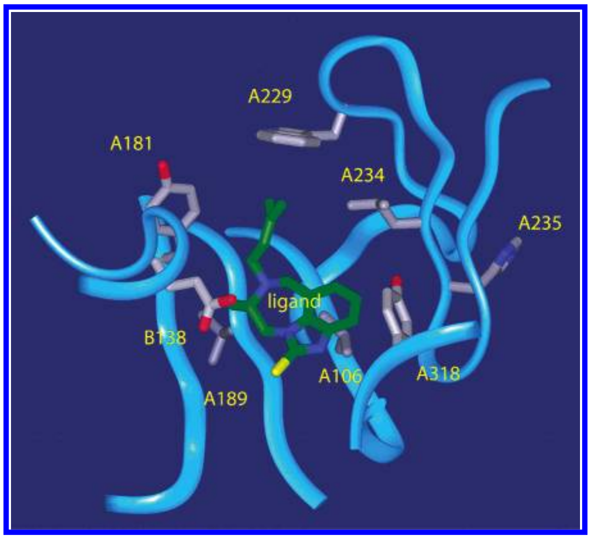

Figure 7. The HIV Reverse Transcriptase pocket with a docked ligand displaying the residues selected for the primary splits by CART.

Supporting Information Available: The data set, containing the pIC_{50} values against the wild type and mutant HIV strains and the averaged calculated interaction energies. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hertogs, K.; de Béthune, M. P.; Miller, V.; Ivens, T.; Schel, P.; Van Cauwenberge, A.; Van Den Eynde, C.; Van Gerwen, V.; Azijn, H.; Van Houtte, M.; Peeters, F.; Staszewski, S.; Conant, M.; Bloor, S.; Kemp, S.; Larder, B.; Pauwels, R. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant HIV-1 isolates of patients treated with antiretroviral drugs (PR-RT-Antivirogram). *Antimicrob. Agents Chemother.* **1998**, *42*, 269−276.

(2) Leach, A. R. *Molecular Modelling, Principles and Applications*; Longman: London, 1996.

(3) Hsiou, Y.; Ding, J.; Das, K.; Clark, A. D., Jr.; Boyler, P. L.; Lewi, P.; Janssen, P. A. J.; Kleim, J.; Rosner, M.; Hughes, S. H.; Arnold, E. The Lys103Asn mutation of HIV-1 RT: Novel mechanism of drug resistance. *J. Mol. Biol.* **2001**, *309*, 437−445.

(4) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. G. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, 1984.

(5) Simeonov, V.; Barbieri, P.; Walczak, B.; Massart, D. L.; Tsakovski, S. Environmetric modeling of a potable water data set. *Toxicol. Environ. Chem.* **2001**, *79*, 55−72.

(6) Kuebler, J.; Russell, A. G.; Hakami, A.; Clappier, A.; Van den Bergh, H.; Episode selection for ozone modelling and control strategies analysis on the Swiss Plateau. *Atmos. Environ.* **2002**, *36*, 2817−2830.

(7) Marshall, R. J. The use of classification and regression trees in clinical epidemiology. *J. Clin. Epidemiol.* **2001**, *54*, 603−609.

(8) Fodor, I. K.; Kamath, C.; Dimension reduction and the classification of bent double galaxies. *Comput. Stat. Data Anal.* **2002**, *41*, 91−122.

(9) Put, R.; Perrin, C.; Questier, F.; Coomans, D.; Massart, D. L.; Vander Heyden, Y. Classification and regression tree analysis for molecular descriptor selection and retention prediction in chromatographic quantitative structure-retention relationship studies. *J. Chromatogr. A* **2003**, *998*, 261−276.

(10) De'Ath, G.; Fabricius, K. E. Classification and regression trees: a powerful yet simple technique for the analysis of complex ecological data. *Ecology* **2000**, *81*, 3178−3192.

(11) Chou, P. A. Optimal partitioning for classification and regression trees. *IEEE Trans. Pattern Anal. Machine Learning* **1991**, *13*, 340−354.

(12) Bose, S. Multilayer statistical Classifiers. *Comput. Stat. Data Anal.* **2003**, *42*, 685−701.

(13) Mehta, M.; Rissanen, J.; Agrawal, R. MDL-based decision tree pruning. Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD '95)

(14) Cappelli, C.; Mola, F.; Siciliano, R. A statistical approach to growing a reliable honest tree. *Comput. Stat. Data Anal.* **2002**, *38*, 285−299.

(15) Daeyaert, F.; de Jonge, M.; Heeres, J.; Koymans, L.; Lewi, P.; Vinkers, M. H.; Janssen, P. A. J. A pharmacophore docking algorithm and its

application to the cross-docking of 18 HIV-NNTI's in their binding pockets. *PROTEINS* **2003**, in press.

(16) Lewi, P. J.; de Jonge, M.; Daeyaert, F.; Koymans, L.; Vinkers, M.; Heeres, J.; Janssen, P. A. J.; Arnold, E.; Das, K.; Clark, A. D. Jr.; Hughes, S. H.; Boyer, P. L.; de Béthune, M.-P.; Pauwels, R.; Andries, K.; Kukla, M.; Ludovici, D.; De Corte, B.; Kavash, R.; Ho, C. On the detection of multiple-binding modes of ligands to proteins, from biological, structural, and modelling data. *J. Comput.-Aided Mol. Design*, in press.

(17) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parametrization and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(18) Hertogs, K.; de Béthune, M.-P.; Miller, V.; Ivens, T.; Schel, P.; Van Cauwenberge, A.; Van Den Eynde, C.; Van Gerwen, V.; Azijn, H.; Van Houtte, M.; Peeters, F.; Staszewski, S.; Conant, M.; Bloor, S.; Kemp, S.; Larder, B.; Pauwels. R. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant HIV-1 isolates of patients treated with antiretroviral drugs (PR-RT-Antivirogram). *Antimicrob. Agents Chemother.* **1998**, *42*, 269−276.

(19) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46*, 2765−2773.

(20) Hsiou, Y.; Das, K.; Ding, J.; Clark, A. D., Jr.; Kleim, J.-P.; Rosner, M.; Winkler, I.; Riess, G.; Hughes, S. H.; Arnold, E.; Structures of Tyr188Leu mutant and wild-type HIV-1 reverse transcriptase complexed with nonnucleoside inhibitor HBY 097: inhibitor flexibility is a useful design feature for reducing drug resistance. *J. Mol. Biol.* **1998**, *284*, 313−323.

(21) Hopkins, A. L.; Ren, J.; Tanaka, H.; Baba, M.; Okamato, M.; Stuart, D. I.; Stammers, D. K. Design of MKC-442 (Emivirine) analogues with improved activity against drug-resistant HIV mutants. *J. Med. Chem.* **1999**, *42*, 4500−4505.

(22) Ren, J.; Diprose, J.; Warren, J.; Esnouf, R. M.; Bird, L. E.; Ikemizu, S.; Slater, M.; Milton, J.; Balzarini, J.; Stuart, D. I.; Stammers, D. K. Phenerhylthiazolylthiourea (PETT) nonnucleoside inhibitors of HIV-1 and HIV-2 reverse transcriptases. *J. Biol. Chem.* **2000**, *275*, 5633−5639.

(23) Ren, J.; Milton, J.; Weaver, K. L.; Short, S. A.; Stuart, D. I.; Stammers, D. K. Structural basis for the resilience of Efavirenz (DMP-266) to drug resistance mutations in HIV-1 reverse transcriptase. *Structure* **2000**, *8*, 1089−1094.

(24) Ding, J.; Das, K.; Moereels, H.; Koymans, L.; Andries, K.; Janssen, P. A.; Hughes, S. H.; Arnold, E. Structure of HIV-1 RT/TIBO R86183 complex reveals similarity in the binding of diverse nonnucleoside inhibitors. *Nat. Struct. Biol.* **1995**, *2*, 407−415.

(25) Ren, J.; Esnouf, R.; Hopkins, A.; Ross, C.; Jones, Y.; Stammers, D.; Stuart, D. The structure of HIV-1 reverse transcriptase complexed with 9-chloro-TIBO: Lessons for inhibitor design. *Structure* **1995**, *3*, 915−926.

(26) Hopkins, A. L.; Ren, J.; Esnouf, R. M.; Willcox, B. E.; Jones, E. Y.; Ross, C.; Miyasaka, T.; Walker, R. T.; Tanaka, H.; Stammers, D. K.; Stuart, D. I. Complexes of HIV-1 reverse transcriptase with inhibitors of the HEPT series reveal conformational changes relevant to the design of potent nonnucleoside inhibitors. *J. Med. Chem.* **1996**, *39*, 1589−1600.

(27) Ren, J.; Esnouf, R. M.; Hopkins, A. L.; Warren, J.; Balzarini, J.; Stuart, D. I.; Stammers, D. K. Crystal structures of HIV-1 reverse transcriptase in complex with carboxanilide derivatives. *Biochemistry* **1998**, *37*, 14394−14403.

(28) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High-resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nature Struct. Biol.* **1995**, *2*, 293−302.

CI034170H