

Use of Classification Regression Tree in Predicting Oral Absorption in Humans

Jane P. F. Bai,^{*,†} Andrey Utis,[#] Gordon Crippen,[‡] Han-Dan He,[§] Volker Fischer,[§] Robert Tullman,[§] He-Qun Yin,[§] Cheng-Pang Hsu,^{||} Lan Jiang,[⊥] and Kin-Kai Hwang[⊥]

ZyxBio, LLC, P.O. Box 2255, Hudson, Ohio 44236, ZyxBio, LLC, 11,000 Cedar Avenue, Cleveland, Ohio 4106, College of Pharmacy, University of Michigan, Ann Arbor, Michigan, Novartis Pharmaceuticals, East Hanover, New Jersey, Johnson & Johnson Pharmaceutical Research & Development, LLC, and Aventis Pharmaceuticals, Bridgewater, New Jersey

Received March 31, 2004

The purpose of this study is to explore the use of classification regression trees (CART) in predicting, in the dose-independent range, the fraction dose absorbed in humans. Since the results from clinical formulations in humans were used for training the model, a hypothetical state of drug molecules already dissolved in the intestinal fluid was adopted. Therefore, the molecular attributes affecting dissolution were not considered in the model. As a result, the model projects the highest achievable fraction dose absorbed, providing a reference point for manipulating the formulations or solid states to optimize oral clinical efficacy. A set of approximately 1260 structures and their human oral pharmacokinetic data, including bioavailability and/or absorption and/or radio-labeled studies, were used, with 899 compounds as the training set and 362 the test set. The numerical range of the fraction dose absorbed, 0 to 1, was divided into 6 classes with each class having a size of approximately 0.16. A set of 28 structural descriptors was used for modeling oral absorption without considering active transport. Then, a separate branch was created for modeling oral absorption involving active transport. The AAE of the training set was 0.12 and those of five test sets ranged from 0.17 to 0.2. In terms of classification, two test sets of unpublished, proprietary compounds showed 79% to 86% prediction when the predicted values fallen within \pm one class of real values were considered predicted. Overall, the computational errors from all the test sets of diverse structures were similar and reasonably acceptable. As compared to artificial membranes for ranking drug absorption potential, prediction by the CART model is considered fast and reasonably accurate for accelerating drug discovery. One can not only improve continuously the accuracy of CART computations by expanding the chemical space of the training set but also calculate the statistical errors associated with individual decision paths resulting from the training set to determine whether to accept individual computations of any test sets.

INTRODUCTION

According to the recent statistics, it takes approximately 14 years and \$850 million to discover, develop, and obtain the FDA approval of a new drug before it can be commercialized.¹ Obviously, there is a tremendous need for high-throughput screening methods to accelerate drug discovery and to cost-effectively produce new cutting-edge medicines. Since ADME (absorption, distribution, metabolism, and elimination) properties are important parameters in lead identifications, using in silico methods to search for drug candidates with good ADME properties has attracted a lot of interest in the pharmaceutical industry.^{2,3} If reasonably accurate and statistically robotic, these methods will help identify ideal drug candidates while drastically shortening the time and reducing the expenses spent in the research and development of new drugs.

Though there are quite a few in vitro high-throughput screening methods, such as Caco-2 monolayers and artificial membranes, available for identifying lead compounds with good oral absorption,⁴ these methods need substantial resources for synthesis and laboratory studies. On the contrary, in silico computations, if reasonably accurate, will provide the chemists a powerful means to create better new drug candidates even without any laboratory syntheses. To properly use the in silico methods, one has to thoroughly understand the shortcomings and strengths of in silico methods and apply them with scientific judgments. In silico models are unlikely to be quantitatively precise for many reasons. First of all, it is not possible to fully capture the true molecular attributes of drug molecules involved in physiological absorption, which determine how molecules move from the hydrated state in the intestine into the nonhydrated state in the lipid bilayers, let alone the integrated results of the fraction dose absorbed. One can use our human perceptions to determine what structural attributes favor drug absorption, but it is unlikely that in silico computation will be absolutely flawless. It is important to understand that all prediction methods draw some sort of statistical conclusions based on the molecular attributes given and the chemical space of the training set. Moreover, there are no perfect data-

* Corresponding author phone/fax: (330)528-1461; e-mail: jbai@zyxbio.com.

[†] ZyxBio, LLC, Hudson, OH.

[‡] University of Michigan.

[§] Novartis Pharmaceuticals.

^{||} Johnson & Johnson Pharmaceutical Research & Development.

[⊥] Aventis Pharmaceuticals.

[#] ZyxBio, LLC, Cleveland, OH.

mining methods even if a perfect set of structural descriptors is used. Having these in mind, one should still have fair confidence in the statistical acceptability of any *in silico* methods if the models offer consistent accuracy and statistical indices to allow proper decision-makings.

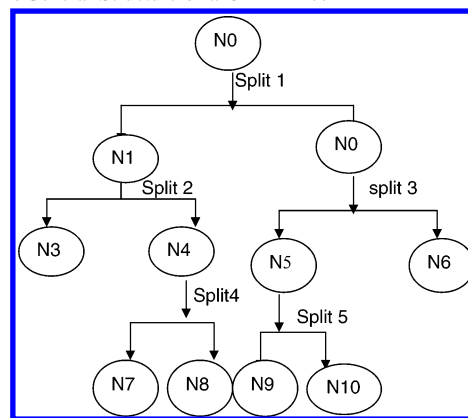
Various data mining and artificial intelligent methods including principal component analysis, fragment method, neural networks, and genetic algorithms have been reported for the development of *in silico* computational methods.^{3,5-7} Though recursive classification regression trees (CART) have been used in predicting diseases from a battery of clinical symptoms,^{8,9} it has not been reported for *in silico* ADME predictions. Compared to other data mining methods, CART has the advantage of requiring only a smaller number of data. Since a large number of pharmacokinetic data are not readily available, CART is considered ideal for ADME applications. Moreover, CART allows one to track down the decision path for a specific computation and to determine whether to adopt the computation based on the error associated with the specific decision path resulting from the training set. This report will discuss the use of CART in predicting the fraction dose absorbed in humans.

In light of intestinal absorption involving multiple processes including several active transport systems (di-/tripeptide and amino acid transporters), efflux by P-glycoprotein, passive transport, paracellular transport, and bile salts acting as absorption enhancers,¹⁰ CART is considered ideal for taking into considerations all these physiological processes, either in series or in parallel. These transport processes are the components contributing to the integrated results of the fraction dose absorbed in humans and were parametrized for developing a CART model.

A data set of approximately 1260 drugs/drug candidates and their human pharmacokinetic data, which was compiled through extensive literature reviews, was used.¹¹ Since all the structures with their published human oral data were collected without any restrictions on pharmacological effects, this data set is considered structurally diverse and statistically robotic to render any computational methods useful for unseen new structures. Only those data from the dose-independent absorption range, from aqueous solution, and from clinical formulations, in which dissolution was not a rate-limiting factor or no controlled release technologies or emulsions were used, were adopted. As a result, the model adopted a hypothetical state of drug molecules already dissolved in the aqueous phase and predicted the highest absorption potential in the dose-independent range, providing additional advantage of serving as a reference for optimizing oral efficacy.

For absorption involving passive transport component and apical efflux, the important well-known molecular attributes,² including computed LogP (octanol/water), the numbers of hydrogen bond donors (HBDs) and acceptors (HBAs), molecular weight, and polar surface area, were used. Furthermore, intramolecular hydrogen bonding was considered. During the absorption process, it is conceivable that drug molecules would have to desolvate themselves from the aqueous phase before partitioning into the lipid membrane.¹² In the intestinal fluid, drug molecules would be fully solvated by water molecules via hydrogen bonds, while in the lipid membrane, drug molecules would minimize the exposure of any hydrogen bond donors/acceptors or other

Scheme 1. General Structure of a CART Tree^a



^a N: node.

polar structural components in order to permeate through the cell membrane. To simulate intramolecular hydrogen bonding, a graph algorithm of nodes (atoms) and edges (bonds) was created to capture intramolecular hydrogen bonding by considering the structural flexibility and steric hindrance between each pair of donor and acceptor. Intramolecular hydrogen-bonding pairs of several compounds identified by such graphical algorithm were compared with those identified by MOE.¹³ It was concluded that the graph algorithm created for identifying intramolecular hydrogen bonding was comprehensive and useful. For the active transport component, the structures of published substrates were analyzed and used to simulate transport via di-/tripeptide and amino acid transport systems.¹⁴

For the sake of comparison, a set of 39 compounds and their real human absorption data versus projected values calculated from artificial membrane permeability data were included in this report to demonstrate that, as compared to *in vitro* methods, the *in silico* methods are fast and reasonably accurate and deserve continuing scientific efforts.

METHODS

Classification and Regression Trees. CART is a powerful data mining-method suitable for knowledge discovery from a small set of data while conveniently accepting both categorical and numerical data. One advantage is that it makes no underlying assumptions regarding the distribution of the values of the predictors. The construction of a tree involves growing the tree to facilitate discoveries, stopping tree growth, and then pruning the tree to offer statistical protection so that the resulting tree from the training set does not have a biased form.¹⁵ The decision tree asks a binary question at each node, yes or no, concerning a specific feature, and partitions the data into 2 subsets resulting in 2 child nodes while minimizing the mean "impurity" of the two partitions at each node in the tree, as summarized in Scheme 1.

For our case, each structural descriptor is a node where the data were split into two subsets according to a certain cutoff with one subset having the frequency of the occurrence of a specific descriptor greater than the cutoff and the other subset less than the cutoff. The cutoff of each structural descriptor was determined statistically according to the distribution of its numerical values in the training set. From the training set, the CART model asked binary questions

regarding individual structural descriptors and determined statistically the suitable cutoffs for individual descriptors for partitioning the data into classes.

The process of partitioning and minimizing the impurity and error of partition continues recursively until some stop criteria are met. The tree is then pruned back according to the complexity parameter, which is used to determine whether the accuracy of additional split adds to the entire tree warrant the split. The leaf node is the terminal node without any further child node.

The general structure of a decision tree is determined by the training set and structural descriptors used. If the tree is statistically well built and not overtrained, then the prediction error from the test set should be close to that for the training set. Of course, this applies to any data-mining methods in general.

The CART models were built using the published methods and their corresponding software.^{16,17} Both classical CART and modified CART which is capable of nonlinear regression were tested in earlier experiments. The latter is able to provide continuous numerical outputs, while the former discrete numbers representing individual classes. Our findings indicated that both methods had similar errors, with the latter showing minimally better prediction. Hence, it was decided to use the classical CART method with classes for our model since it is easier to implement.

Classification of Fraction Dose Absorbed. In this reported CART model, the numerical range of the fraction dose absorbed was divided into six classes, 0–0.19, 0.2–0.31, 0.32–0.43, 0.44–0.59, 0.6–0.75, and 0.76–1. The predicted data were reported as the median value of a class range. Therefore, it is expected that the intrinsic error, on the average, would be 0.07 (half of a class size). Unfortunately, the classification aforementioned might also cause other inherent computation deviations and errors. That is, the predicted values may fluctuate due to the possibilities of the predicted value right on the borderline of two adjacent classes and due to the nature of numbers stored by the computer. Some real numbers may be stored as numbers with indefinite digits after the decimal points. Consequently, if some predicted value is right on the borderline between classes 3 and 4, it could be reported by the computer to be either in class 3 or in class 4. Both classes are deemed correct by the computer but would result in the predicted data fluctuating. By the same token, some training data could also have equal possibilities to be in any one of two adjacent classes, thereby contributing to the statistical deviations or errors in predictions. Certainly, manipulating of the training data to avoid such data fluctuation without introducing additional data errors can be done. In summary, one has to exercise one's judgments and understandings of the pros and cons involved in *in silico* computations and data-mining methods to make the best use of them for accelerating drug discovery.

Structural Descriptors. There are 28 structural descriptors used for modeling absorption involving passive transport and apical efflux. In essence, the structural descriptors were chosen to reflect the involvement of many transport processes in oral absorption including para-cellular, trans-cellular passive transport, and P-glycoprotein-mediated efflux. For paracellular transport, molecular weight is the determining factor. Notably, lipophilic cations with multiple cyclic rings were good P-glycoprotein substrates.¹⁸

A separate branch was built for those structures of which absorption involved active transport, passive transport, and apical efflux. The structural requirements for entering this branch for the di-/tripeptide type molecules included one peptide bond or two peptide bonds and at least a free alpha-amino or a carboxylic acid or a carboxylic ester group.¹⁴ For active transport via amino acid carriers, the existence of both amino group (alpha or beta or gamma position) and carboxylic acid was structurally required.¹⁴ The structural requirements described above for active transport along with the 28 descriptors were used for building the active transport branch. Only those identified to fit the structural characteristics of carrier-mediated substrates will enter the specific carrier-mediated branch. In the database, there are approximately 200 compounds of beta-lactam antibiotics, ACE inhibitors, amino acids, and di-/tripeptides analogues. Random selection of two-thirds of the structures was for the training set and the rest for the test set. The active transport contributions were predetermined using some well-characterized substrates.¹⁴ The contribution from active transport was 0.5 for those having alpha-amino and carboxylic acid groups, 0.3 for those with amino group in the beta position, 0.2 those without amino group in the alpha or beta position while containing a free carboxylic acid, 0.1 for those without amino group but with a carboxylic ester, and 0 for those with a carboxylic acid group at the alpha position like enalaprilat. Though somewhat subjectively assigned, these numerical contributions were determined according to the results published for cefuroxime axetil, enalapril, enalaprilat, cephalexin, and captopril, L-dopa, alpha-methyldopa.^{14,19–22}

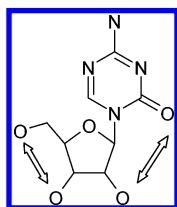
For the selection of the 28 structural descriptors, we referred to the literature reporting the transport mechanisms and the factors affecting lipophilicity, polarity, and absorption across the biological membrane. LogP and polar surface area of each structure were computed according to the published atom-based calculations.^{23–27} These well-known nonproprietary structural descriptors have been reported to affect drug absorption across the biological membrane.^{23–27} As stated in the Introduction, intramolecular hydrogen bonding is likely to increase absorption through biological membrane. During absorption, drug molecules would need structural transformation in order to permeate through the biological membrane; therefore, structural flexibility due to rotatable single bond or structural rigidity due to double bond in the noncyclic structural component would likely affect drug absorption. Molecular weight reflects the size for creating a cavity in the biological membrane for penetration. Aromatic rings can interact with the biological membrane via charge transfer interactions due the electron density inside the aromatic rings, thereby affecting absorption across the biological membrane in which lipids, proteins, and enzymes are embedded. We conducted linear regressions between the frequency of occurrence or the computed numerical value of each descriptor and the fraction dose absorbed of the whole data set. The correlation coefficients of some structural descriptors from the regression analysis are listed in Table 1. Notably, the quaternary amine and the number of sulfate and phosphate groups have low correlation coefficients in the whole set of 1260 data because of their low frequency of occurrence in the whole set. However, they had a higher correlation when compounds with higher absorption were excluded and had the highest correlation if only those with

absorption less than 0.2 were used. For example, the correlation coefficient for the quaternary amine reached the 0.09 for those with absorption less than 0.2. These results were not surprising since ionization impedes absorption across the biological membranes, and compounds with low absorption have a higher frequency of occurrence of those structural descriptors that impede membrane transport.

The CART method offers the advantage of placing more weights of certain structural descriptors in a class where these structural descriptors have higher frequency of occurrence. This is why we decided to test the CART model in computing oral absorption since it attempts to reach the minimum statistical errors within individual classes.

According to the literature, the H-bond acceptor is in general an atom that has high electron density for interacting with acidic hydrogen.²⁶ The numbers of HBA and HBD were determined according to the published definitions.^{2,26} The hydrogen bond donors include $-NH_2$, $-OH$, and $-NHX$ where X is any atom other than hydrogen and hydrogen bond acceptors the nitrogen atom of $-NH_2$, the oxygen atom of $-OH$, the carboxyl group, and so on.²⁶ The number of intramolecular hydrogen bonding was determined using a hypothetical intramolecular interaction algorithm, which pairs HBAs and HBDs with extensive considerations of steric hindrance between any pairs of HBA and HBD.²⁷ In principle, the more single rotatable bonds are between a pair of HBA and HBD, the more likely intramolecular H-bonding will occur. Formation of six-member rings is more favorable than the formation of five-member rings and was given higher priority. With the aid of such graph algorithm, identifications of intramolecular H-bonding seemed to agree, to a great extent, with those through force field computations.¹³

As shown below, the number of intramolecular H-bonds identified for azacitidine agreed with the results of some conformers of low energies in the gas-state identified using MOE.¹³



More such comparisons were done with additional drugs, and the results showed that our graph algorithm for identifying intramolecular bonding was comparable to MOE in terms of the number of intramolecular hydrogen bondings in individual drugs and the individual atoms involved.

To enable digitizing molecular structures for programming, two-dimensional chemical structures were converted to SMILES (Simplified Molecular Input Line Entry System), which were generated using mole2.smi.²⁸ SMARTS notation published on the Daylight Chemical information Inc. was adopted for reading in structural descriptors. The frequency of occurrence for each structural descriptor was then determined.

Training and Test Sets Used in CART Model. In this study, we aimed at developing a model for predicting absorption in the dose-independent or concentration-independent range. A database of approximately 1260 drugs and drug

Table 1. List of Some Nonproprietary Structural Descriptors Used

descriptors	r ² ^a (whole data set)	r ² ^a (absorption 0–0.19)
computed Log P (octanol/water)	0.067	
number of quaternary amine functional group	0.005	0.09
number of hydrogen bond donors plus acceptors	0.15	
number of computed intramolecular hydrogen bonds	0.09	
molecular weight	0.18	
computed polar surface area	0.29	
number of double bond in the noncyclic structural component	0.06	
number of aromatic rings	0.086	
number of rotatable single bonds	0.02	
number of carboxylate functional group	0.02	
number of sulfate functional group	0.001	0.046
number of phosphate functional group	0.002	0.065
presence of metal in the structure	0.001	0.037

^a The correlation coefficient obtained from the linear regression of the actual absorption data versus the frequency of occurrence or numerical value (for the case of molecular weight) of a descriptor.

Table 2. Distribution of Oral Absorption of Approximately 1260 Structures in OraSpotter Database

absorption range	percentage (%)
structures with absorption ≤ 0.4	24
$0.4 <$ structures with absorption < 0.7	12.2
structures with absorption ≥ 0.7	63.4

candidates as well as their human pharmacokinetic data was used.¹¹ Only those data from micronized solids were adopted if absorption was dissolution-dependent. Since the data were randomly extracted from the literature and not limited to drug/drug candidates for a specific disease, any prediction methods developed from this database would presumably be statistically robotic to render acceptable prediction accuracy for unseen chemical structures. The distribution of the values of oral fraction dose absorbed in humans in this database is summarized in Table 2. One might criticize that the distribution of the values of fraction dose absorbed is skewed toward high absorption. Being aware of this criticism, earlier experiments chose an equal number of compounds from high, medium, and low absorption, built a CART model, and tested with the remaining data set. We then compared the predictions with those from a CART model with a training set consisting of 70% of data in the high absorption range. The results indicated that the distribution of the training data did not make any significant differences in terms of the average absolute errors observed in the test sets.

With the relief of CART being able to handle skewed data, the database was randomly divided into a training set of 899 compounds and a test set of 362 compounds (test set A) for building and testing a CART model. The random split of the database into a training set and a test set was experimented a few times to determine whether distinct training sets have any influences on the resulted trees and on the accuracy of prediction. The average absolute errors (AAE) was estimated as reported in the literature,³ and the errors from different random splits of the database were compared. Furthermore, the percents of correct classification versus those of misclassification were compared. Since the predicted values were discrete numbers representing the medians of individual classes, AAE was considered more appropriate

to show the prediction quality of the model than the correlation coefficients between the predicted and real data. Similar individual AAEs were observed among individual training sets, so were the prediction errors from individual test sets. From these experiments, it was concluded that the structural descriptors used were not biased. Importantly, these experiments also served to validate the structural descriptors.

Further Validation with Independent Proprietary Test Sets. In addition to the test A aforementioned, further validations were carried out using a small set of literature data and two other independent proprietary sets of data. Test B of ritonavir analogues²⁹ and test set C consisting of 65 structures and their human and rat data was compiled from the literature. Test set D consisted of 90 proprietary structures and their in vivo rat, mouse, monkey, and dog absorption data using radiolabeled compounds dissolved in solutions. Test set E consisted of 37 proprietary structures of antifungal, antihistamine, antiemetics, CNS, and anticancer drugs. The fraction dose absorbed data of test set D and E were obtained by a mass balance of radiolabeled studies using excretions in urine and feces. The underlying assumption made in radiolabeled studies was that there was no significant intestinal metabolisms during absorption and no enterohepatic recycling. These diverse test sets differed greatly in their physicochemical properties with water solubility ranging from very soluble to insoluble.

RESULTS

Impact of Classification Schemes. Earlier experiments tested the impact of two-tier classification. A cutoff of 0.5 or 0.6 or 0.75 was chosen, and the numerical range of fraction dose absorbed, 0 to 1, was divided into two classes of above and below the cutoff. Then the class of low absorption was further classified into 6 subclasses. The purpose of using this two-tier classification was to have better prediction for low absorption in the hope of providing a means to accurately eliminate structures, which will not be orally active. Disappointedly, the two-tier classification did not offer any advantages of lowering the prediction errors. The AAE in the training set from the two-tier classification was slightly higher than that from one-tier classification. Clearly, once a compound of low absorption was misclassified in the high absorption class, it would never be reclassified correctly in the low absorption class, and vice versa. Therefore, one-tier classification was chosen for building model.

Impacts of Data Distribution in the Training Sets and Different Training Sets. Since OraSpotter database has a distribution skewed toward high absorption range (Table 2), it is important to determine whether the biased data distribution in the training set would cause biased prediction. A few experiments were performed using an even data distribution by randomly choosing 300 structures from high, medium and low absorption and tested the CART model on the rest of structures. The result showed that the errors resulted from an even distribution of training data were similar to those resulted from a random selected set of which data were skewed toward high absorption. In summary, data distribution of the training set does not affect how CART performs. This seems to agree to the general belief that CART can deal with data with skewed distributions.

To test whether the model is biased by a specific training set, earlier experiments examined the models resulting from

Table 3. Distribution of Real versus Predicted of 362 Structures in Test Set A^a

predicted	real					
	0–0.19	0.2–0.31	0.32–0.43	0.44–0.59	0.6–0.75	0.76–1
0–0.19	[60]	(6)	3	4	3	1
0.2–0.31	(1)	[1]	(2)	(1)	1	5
0.32–0.43	0	(0)	[1]	(0)	2	2
0.44–0.59	1	3	(1)	[2]	(3)	11
0.6–0.75	2	2	1	(3)	[1]	(6)
0.76–1	5	4	6	15	(32)	[172]

^a All data are human absorption data collected from the literature. Absorption from 0 to 1 is divided into 6 classes as shown in the table.

different random training sets and the results showed similar average absolute errors (AAE) from any random training sets tested.

Results of the Training Set and Test Set A. For the final CART model, the training set of 899 structures had an AAE of 0.12; in other words, the real value was, on the average, the predicted value \pm 0.12. This prediction error was considered reasonable in light of the fact that the average intrinsic error was 0.075 on account of the median value of a class being reported for each prediction.

For test set A of 362 structures, the distribution of real versus predicted values of test set A is detailed in Table 3. The bracketed data in the table represent the number of structures of which prediction is exactly in the right class, those in the parentheses the number of structures of which predicted data are one class higher or lower than the real data.

Test A consisting of 362 structures had an AAE was 0.169, which is, as expected, higher than but reasonably close to that of the training set. More than one-half of the structures, 65%, in test set A was classified in the right class, and 80.4% predicted within one class error (\pm one class). For structures with absorption less than 0.16, 1.9% was predicted to have absorption higher than 0.6. For structures with absorption higher 0.82, 0.2% was predicted to have absorption lower than 0.19.

In light of pharmacokinetic variations due to genetic differences, gender differences, and different ethnic backgrounds, the resulting AAE of test set A seemed to support CART as an acceptable data mining method. In ideal predictions, Table 3 should only have data distributed along its diagonal. Though the prediction of test set A was not perfect, there is a high concentration of data distributed along the diagonal of Table 3, suggesting that the CART model is reasonable.

Test Set B and the Advantages of Revealing the Decision Path Leading to a Computed Result and Individual Prediction Errors Associated with Individual Decision Paths. Table 3 summarizes the predictions for test set B consisting of ritonavir analogues and their real rat data.²⁹ The rat data with little first pass metabolisms were adopted from the literature. Though ritonavir and some of its analogues were not predicted, some of its analogues were predicted accurately.

CART offers the advantage of revealing the decision path for each computation, which is the path taken by the model for reaching a specific predicted value for a specific structure. One can trace backward from the end leaf node of a

Table 4. Results of Ritonavir and Its Analogs in Test Set B^a

ritonavir and its analogues	computed results	individual computation paths	actual data
analogue #119	0.52	2a,1b,3b,1b,4b	0.47
analogue #141	0.52	2a,1b,3b,1b,4b	1.
analogue #157	0.52	2a,1b,3b,1b,4b	0.76
analogue #130	0.88	2a,1a,3b,23b,13b,22b,7b,2a,25b,22b	1
analogue #134	0.88	2a,1a,3b,23b,13b,22b,7b,2a,25b,22b	1
analogue #131	0.26	2a,1b,3a,22b,18b,26a,4b	0.7
analogue #132	0.26	2a,1b,3a,22b,18b,26a,4b	0.57
analogue #135	0.26	2a,1b,3a,22b,18b,26a,4b	0.58
analogue #137	0.26	2a,1b,3a,22b,18b,26a,4b	0.6
analogue #150	0.26	2a,1b,3a,22b,18b,26a,4b	0.23
ritonavir	0.26	2a,1b,3a,22b,18b,26a,4b	0.76

^a First column: list of ritonavir and its analogues which are numbered as shown in the ref 29; second column: the computed results; third column: individual decision paths adopted by the software for individual compounds; fourth column: real data.

computed result to the starting node to obtain the decision path. As shown in Table 4, ritonavir analogues of #119, #141, and #157 shared the same decision path of 2a,1b,3b,1b,4b chosen by the CART model, while other ritonavir analogues shared other paths. Those sharing the same decision path have the same numerical output. Therefore, one can examine individual chemical structures and their decision paths as well as computed results for decision-makings.

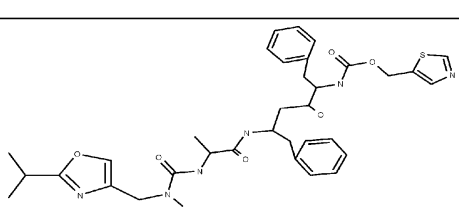
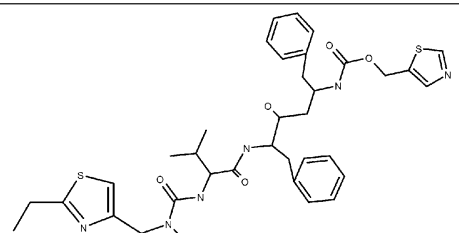
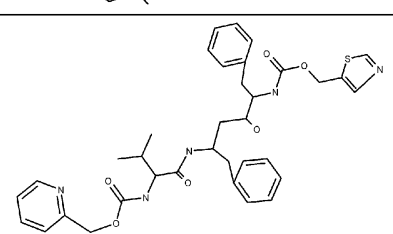
Furthermore, one can even calculate individual errors associated with individual decision paths from a large training set and then use the errors associated with individual decision paths to judge the predicted results for any given test sets. Let us use ritonavir and its analogues to illustrate how one can use the statistical errors associated with individual

decision paths to determine what predictions to adopt. From the training set, the decision path of 2a,1b,3b,1b,4b had an AAE of 0.03 and 100% correct classification, that of 2a,-1b,3a,22b,18b,26a,4b had an AAE of 0.19 and 67% correct classification, and that of 2a,1a,3b,23b,13b,22b,7b,2a,25b,-22b had an AAE was 0.14 and 80% correct classification. In short, the first decision path was most accurate, followed by the third, and then the second. Consequently, one could have the highest confidence in the predictions for ritonavir analogues # 119 and #157 since the error associated with their decision path is 0.03, which is much smaller than 0.12 from the whole training set. By the same rationale, the computations for analogues #130 and # 134 could also be adopted. As for ritonavir and its analogues #131, #132, #135, #137, and #159, one would be less confident since the error associated with their decision path was much higher than that of the training set. From Table 4, it is obvious that the overall AAE for analogues #130 and #134 was 0.04, that for analogues #119 and #157 was 0.145, and that for ritonavir and analogues #131, #132, #135, #137, and #159 was 0.32. The errors associated with individual decision paths taken for ritonavir and its analogues agreed with the errors calculated from their real and predicted data. Importantly, one can re-train the model with ritonavir and its analogues of which AAE was larger to make the model predict these type of compounds better.

The structures of ritonavir analogues whose prediction was correct by the model are compared in Table 5. These analogues sharing similar structural characteristics took different decision paths in the model.

Test Set C. Test set C of 67 structures and their rat and human oral absorption data were gathered from the literature.

Table 5. Chemical Structures of Ritonavir Analogs^a

Name	Structure	Computed data	Real data
#150		0.26	0.23
#130		0.88	1
#119		0.52	0.47

^a Rat data were taken from ref 29.

Table 6. Distribution of Real versus Predicted Data of 67 Structures in Test Set C^a

predicted	real					
	0–0.19	0.2–0.31	0.32–0.43	0.44–0.59	0.6–0.75	0.76–1
0–0.19	[2]	(0)	1	0	1	1
0.2–0.31	(0)	[0]	(1)	1	1	0
0.32–0.43	0	(0)	[0]	(0)	0	1
0.44–0.59	0	0	(2)	[1]	(1)	0
0.6–0.75	0	0	1	(1)	[2]	(2)
0.76–1	0	0	0	3	(11)	[34]

^a All data were human absorption data from the literature.

Table 7. Distribution of Real versus Predicted Data of 90 Proprietary Structures in Test Set D^a

predicted	real					
	0–0.19	0.2–0.31	0.32–0.43	0.44–0.59	0.6–0.75	0.76–1
0–0.19	[3]	(1)	0	1	0	2
0.2–0.31	(1)	[0]	(0)	(0)	3	2
0.32–0.43	0	(0)	[1]	(1)	0	1
0.44–0.59	1	0	(0)	[2]	(1)	3
0.6–0.75	1	1	0	(1)	[1]	(5)
0.76–1	0	0	2	6	(8)	[42]

^a The in vivo absorption data were from radio-labeled studies in rats, mice, dogs, and monkeys.

It is assumed that drug absorption in rats and humans is similar. Results of test set C are summarized in Table 6, showing that 85.1% was predicted within ± 1 class, 5.97% underpredicted by more than 3 classes, and 0% overpredicted by 3 classes. Prediction for test set C had an AAE of 0.17, similar to that of test set A.

Test Set D. Test set D consisted of 90 proprietary structures and their absorption data from radio-labeled studies in rats, mice, monkeys, and dogs. Test set D has an AAE of 0.2, slightly larger than those observed for test set A and C. The results are summarized in Table 7, having 54.4% predicted in the right class, 74.4% within \pm one class, 5.5% underpredicted by 3 classes, 4.4% underpredicted by more than 3 classes, 4.4% overpredicted by 3 classes, and 1.1% overpredicted by more than 3 classes.

Test Set E. Test set E consisted of proprietary 37 structures with absorption ranging from 6% to 100%. Their computed, actual data and computation errors are summarized in Table 8. The CART method predicted 67.6% of structure in the right class and 86.4% within one class variation and the resulting AAE was 0.14.

Summary of the Results of All Test Sets. The CART model seems to predict better for high and low absorption but did poorly for intermediate absorption ranging from 0.32 to 0.59. One of the reasons might be that there is a smaller number of compounds with intermediate absorption than those in low and high absorption. Consequently, the CART was not properly built to reflect the molecular characteristics of compounds in this category. To predict better for intermediate absorption, collecting more of this category of compounds is necessary to improve prediction. As the chemical space of the training set expands, it is very likely that the structural descriptors used would need to be modified in order to make the prediction more accurate.

Predictions for Diverse Structures. More diverse structures are listed in Table 9 to demonstrate that CART can

Table 8. Results of the 37 Compounds in Test Set E

proprietary compounds	computed results	real data	errors (computed-real)
cpd # 1	0.88	0.81	0.07
cpd # 2	0.88	0.93	–0.05
cpd # 3	0.88	0.89	–0.01
cpd # 4	0.68	0.78	–0.10
cpd # 5	0.88	0.88	0.0
cpd # 6	0.88	0.90	–0.02
cpd # 7	0.88	0.89	–0.01
cpd # 8	0.88	0.81	0.07
cpd # 9	0.88	0.78	0.1
cpd # 10	0.88	0.71	0.17
cpd # 11	0.88	0.97	–0.09
cpd # 12	0.88	0.90	–0.02
cpd # 13	0.88	0.87	0.01
cpd # 14	0.88	0.47	0.41
cpd # 15	0.88	0.92	–0.04
cpd # 16	0.88	1	–0.12
cpd # 17	0.26	0.85	–0.59
cpd # 18	0.88	0.96	–0.08
cpd # 19	0.88	0.74	0.14
cpd # 20	0.88	0.83	–0.07
cpd # 21	0.88	0.91	–0.03
cpd # 22	0.68	1	–0.32
cpd # 23	0.88	0.88	0.0
cpd # 24	0.88	0.84	0.04
cpd # 25	0.88	0.72	0.16
cpd # 26	0.88	0.06	0.8
cpd # 27	0.88	0.9	–0.02
cpd # 28	0.88	0.93	–0.05
cpd # 29	0.88	0.93	–0.05
cpd # 30	0.88	0.88	0.0
cpd # 31	0.88	0.83	0.05
cpd # 32	0.88	0.69	0.19
cpd # 33	0.88	0.88	0.0
cpd # 34	0.88	0.66	0.22
cpd # 35	0.26	1	–0.74
cpd # 36	0.88	0.69	0.19
cpd # 37	0.52	0.83	–0.31

cover a structurally diverse chemical space offering an acceptable computational accuracy.

In Silico Computation versus in Vitro Artificial Membrane Screenings. To accelerate drug discovery, the pharmaceutical industry is continuously searching for high throughput screening technologies for lead identification. Most recent efforts have been focused on using artificial membrane to identify drug candidates with good oral absorption. Figure 1 compares the predictions of 39 compounds from the CART method (series 1) and artificial membrane studies (series 2). The correlation between the actual absorption data versus the computed values from the CART method was better than that between the actual data versus the predicted data derived from the artificial membrane studies. Better predictions from the CART method could be due to the fact that the presence of some compounds, which are substrates of active transport mechanisms, in the relatively small number of compounds and artificial membranes cannot be applied for compounds involving active transport while the CART method considers that.

Beside the fact of being not being able to handle active transport and energy-dependent efflux by P-glycoprotein, artificial membrane seemed to over-predict oral absorption for some drugs. It is likely that irreversible adsorption of drug molecules to the membrane will cause inaccurate outcome. Though not needing any cell culture like other in

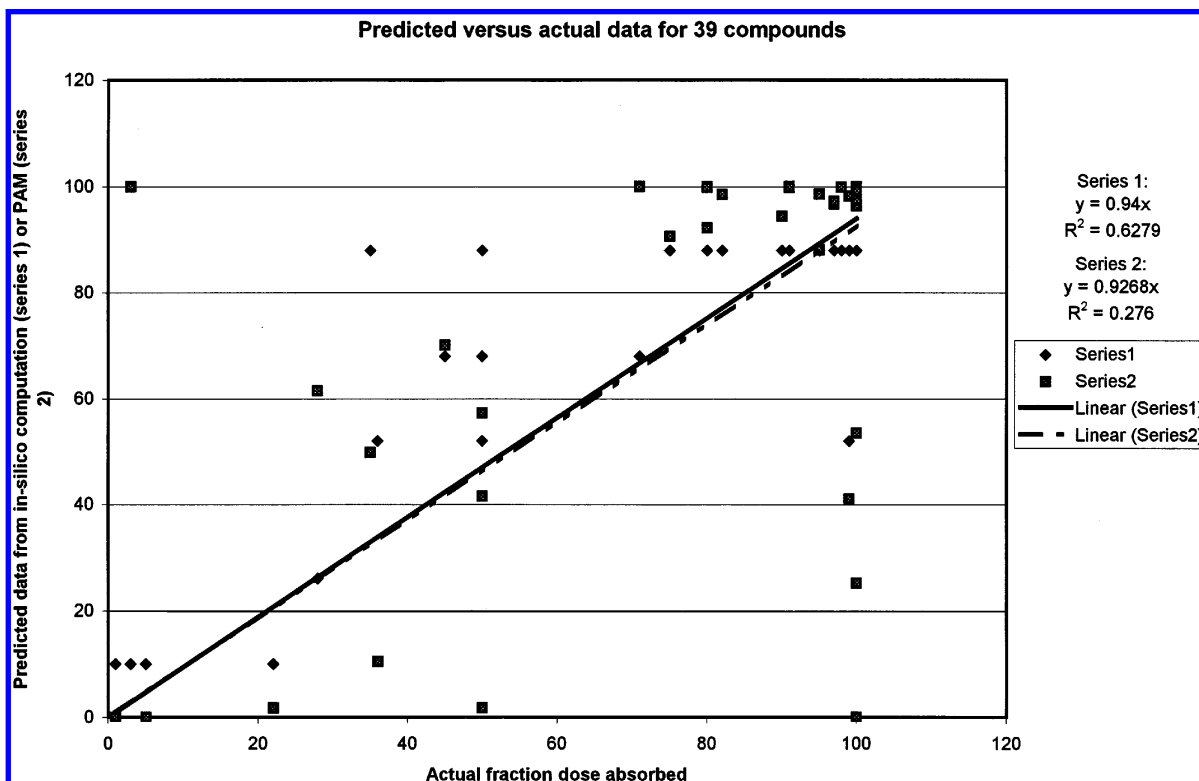
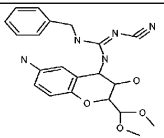
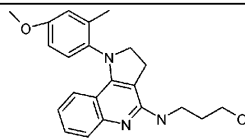
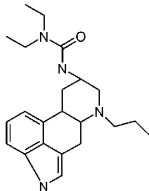
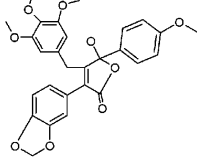
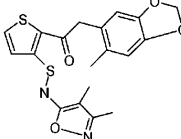


Figure 1. The plot of the actual absorption data of 39 compounds versus their computed values from CART computation (series 1) and their actual absorption data versus the projected absorption values from artificial membrane studies (series 2). Note: CART reports six possible discrete median values of individual predicted classes while the projected data from the artificial membrane studies are continuous.

Table 9. Comparison of Real versus Predicted Data by CART For Some Structures

Name	Structure	Real data (reference)	Computed data
KR-31378		1 (rats, absorption) (ref. 30)	0.68
DBM-819		0.97 (bioavailability; rats) (ref. 31)	0.88
PROTERGURIDE		1 (low dose, rats) (ref. 32)	0.88
PD156707		0.67 (bioavailability; dogs) (Ref. 33)	0.88
TBC11251		0.6 (bioavailability; rats) (Ref. 34)	0.88

vitro cell lines, artificial membrane studies still require chemical synthesis, development of analytical methods, in vitro permeability experiments, and data analysis. In light of the results shown above, it is clear that in silico methods deserve continuous efforts.

CONCLUSION

Judging from the consistent similar acceptable prediction errors for all test sets of very diverse structures, it is concluded CART is useful for in silico prediction of oral absorption. In silico computations might not be less accurate than absorption derived from those in vitro laboratory artificial membrane studies. To improve the prediction accuracy of any in silico models, it is important to have a large training set with very diverse chemical structures so that the model will be statistically robotic and useful for any unseen new structures. Identifying the decision path for each computation resulting from a model and calculation of statistical errors associated with individual decision paths will help the chemists to make the best use of the computed results for modifying their structures.

ACKNOWLEDGMENT

Kind comments from Dr. Wei-Yin Loh (Statistics Department at University of Wisconsin, Madison, Wisconsin) and Mr. Valery Polyakov and Dr. Abdel Laoui of Aventis Pharmaceuticals (Bridgewater, New Jersey) are greatly appreciated.

REFERENCES AND NOTES

- (1) PhRMA Pharmaceutical Industry Profile 2001 (PhRMA Web site).
- (2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Res.* **1997**, 23, 3.
- (3) Gao, H.; Shanmugasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSPR approach. *Pharm. Res.* **2002**, 19, 497.
- (4) Zhu, G.; Jiang, L.; Chen, T. M.; Hwang, K. K. A comparative study of artificial membrane permeability assay for high throughput profiling of drug absorption potential. *Eur. J. Med. Chem.* **2002**, 37, 399.
- (5) McElroy, N. R.; Jurs, P. C. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2001**, 41, 1237.
- (6) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 165.
- (7) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 474.
- (8) Falconer, J. A.; Naughton, B. J.; Dunlop, D. D.; Roth, E. J.; Strasser, D. C.; Sinacore, J. M. Predicting stroke inpatient rehabilitation outcome using a classification tree approach. *Archives Phys. Med. Rehab.* **1994**, 75, 619.
- (9) Mair, J.; Smidt, J.; Kecskutber, P.; Dienstl, F.; Puschendorf, B. A decision tree for the early diagnosis of acute myocardial infarction in nontraumatic chest pain patients at hospital admission. *Chest* **1995**, 108, 1502.
- (10) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-based design: Optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* **2001**, 44, 1313.
- (11) *OraSpotter human pharmacokinetic database*. ZyxBio, LLC: Hudson, OH. (www.zyxbio.com).
- (12) Burton, P. S.; Conradi, R. A.; Hilgers, A. R.; Ho, N. F. H.; Maggiora, L. L. The relationship between peptide structure and transport across epithelial cell monolayers. *J. Controlled Release* **1992**, 19, 87.
- (13) MOE Molecular Operating Environment software. Chemical Computing Group, Inc.: Vancouver, Canada (www.chemcomp.com).
- (14) Bai, P.-F.; Amidon, G. L. Gastrointestinal Transport of Peptide and Protein Drugs and Prodrugs. In *Pharmacokinetics of Drugs*; Welling, P. G., Balant, L. P., Eds.; Springer-Verlag: New York, 1991; pp 189–205.
- (15) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Wadsworth: Belmont, 1984.
- (16) Quinlan, J. R. a. Induction of decision trees. *Machine Learning* **1986**, 1, 81. b. www.cse.unsw.edu.au/quinlan.
- (17) Free software of modified CART capable of nonlinear regression from Dr. Wei-Yin Loh of Statistics Department, University of Wisconsin, Madison, Wisconsin.
- (18) Gros, P.; Talbot, F.; Tang-Wai, D.; Bibi, E.; Kaback, R. Lipophilic cations: A group of model substrates for the multidrug-resistance transport. *Biochemistry* **1992**, 31, 1992.
- (19) Hidalgo, I. J.; Raciassi, S. D. The role of an alpha-amino group on H⁺-dependent transepithelial transport of cephalosporins in Caco-2 cells. *J. Pharm. Pharmacol.* **1999**, 51, 35.
- (20) Sinko, P. J.; Amidon, G. L. Characterization of the oral absorption of beta-lactam antibiotics. I. Cephalosporins: determination of intrinsic membrane absorption parameters in the rat intestine in situ. *Pharm. Res.* **1988**, 5, 645.
- (21) Ruiz-Balaguer, N.; Nacher, A.; Casabo, V. G.; Merino Sanjuan, M. Intestinal transport of cefuroxime axetil in rats: absorption and hydrolysis processes. *Int. J. Pharm.* **2002**, 234, 101.
- (22) Friedman, D. L.; Amidon, G. L. Passive and carrier-mediated intestinal absorption components of two angiotensin-converting enzyme (ACE) inhibitor prodrugs in rats: enalapril and fosinopril. *Pharm. Res.* **1989**, 6, 1043.
- (23) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 868.
- (24) Wang, R.; Fu, Y.; Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **1997**, 37, 615.
- (25) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, 43, 3714.
- (26) Scheiner, S. *Hydrogen bonding*; Oxford University Press: New York, New York, 1977.
- (27) March, J. *Advanced organic chemistry*, 3rd ed.; John Wiley & Sons: New York, 1985.
- (28) mol2.smi. Daylight Chemical information Inc (www.daylight.com): Mission Viejo, California.
- (29) Kempf, D. J.; Sham, H. L.; Marsh, K. C.; Flentge, C. A.; Betebenner, D.; Green, B. E.; McDonald, E.; Vasavanonda, S.; Saldivar, A.; Wideburg, N. E.; Kati, W. M.; Ruiz, L.; Zhao, C.; Fino, L.; Patterson, J.; Molla, A.; Plattner, J. J.; Norbeck, D. W. Discovery of ritonavir, a potent inhibitor of HIV protease with high oral bioavailability and clinical efficacy. *J. Med. Chem.* **1998**, 41, 602.
- (30) Kim, H. J.; Kim, S. H.; Kim, S. O.; Lee, D. H.; Lim, H.; Yoo, S. E.; Lee, M. G. Dose-dependent pharmacokinetics of a new neuroprotective agent for ischemia-reperfusion damage, KR-31378, in rats. *Biopharm. Drug Disp.* **2000**, 21, 279.
- (31) Kim, E. J.; Kim, S. O.; Lee, D. H.; Lim, H.; Lee, M. G. Dose-dependent pharmacokinetics of a new reversible proton pump inhibitor, DBM-819, after intravenous and oral administration to rats: hepatic first-pass effect. *Biopharm. Drug Disp.* **2001**, 22, 119.
- (32) Krause, W.; Hampel, M. Pharmacokinetics of proterguride in rat and cynomolgus monkey. *Xenobiotica* **1988**, 18, 41.
- (33) Patt, W. C.; Edmunds, J. J.; Repine, J. T.; Berryman, K. A.; Reisdorph, B. R.; Lee, C.; Plummer, M. S.; Shahripour, A.; Haleen, S. J.; Keiser, J. A.; Flynn, M. A.; Welch, K. M.; Reynolds, E. E.; Rubin, R.; Tobias, B.; Hallak H.; Doherty A. M. Structure relationships in a series of orally active gamma-hydroxy butenolide endothelin antagonists. *J. Med. Chem.* **1997**, 40, 1063.
- (34) Wu, C.; Chan M. F.; Stavros, F.; Raju, B.; Okun, I.; Mong, S.; Keller, K. M.; Brock, T.; Kogan, T. P.; Dixon, R. A. Discovery of TBC11251, a potent, long acting, orally active endothelin receptor-A selective antagonist. *J. Med. Chem.* **1997**, 40, 1690.

CI040023N