Predicting Human Oral Bioavailability of a Compound: Development of a Novel Quantitative Structure-Bioavailability Relationship

C. Webster Andrews,[1] Lee Bennett,[1,2] and Lawrence X. Yu[1,3,4]

Received January 12, 2000; accepted March 7, 2000

Purpose. The purpose of this investigation was to develop a quantitative structure-bioavailability relationship (QSBR) model for drug discovery and development.

Methods. A database of drugs with human oral bioavailability was assembled in electronic form with structure in SMILES format. Using that database, a stepwise regression procedure was used to link oral bioavailability in humans and substructural fragments in drugs. The regression model was compared with Lipinski's Rule of Five.

Results. The human oral bioavailability database contains 591 compounds. A regression model employing 85 descriptors was built to predict the human oral bioavailability of a compound based on its molecular structure. Compared to Lipinski's Rule of Five, the false negative predictions were reduced from 5% to 3% while the false positive predictions decreased from 78% to 53%. A set of substructural descriptors was identified to show which fragments tend to increase/decrease human oral bioavailability.

Conclusions. A novel quantitative structure-bioavailability relationship (QSBR) was developed. Despite a large degree of experimental error, the model was reasonably predictive and stood up to cross-validation. When compared to Lipinski's Rule of Five, the QSBR model was able to reduce false positive predictions.

KEY WORDS: bioavailability; quantitative structure-bioavailability relationship; Lipinski's Rule of Five.

INTRODUCTION

Recent developments in combinatorial chemistry and high throughput screening techniques have enormously increased the possibility of finding lead compounds (1). However, many lead compounds fail to progress into the clinic because they are lacking appropriate pharmaceutical properties, such as oral bioavailability. There would be many more new drugs than we actually have today if all these lead compounds had desirable biopharmaceutics properties. On the other hand, development of all lead compounds is costly, and cost reduction demands that predictive methods be applied at the preclinical stage to select the best candidate (2). Although oral bioavailability has recently received attention from chemists, no quantitative guideline exists regarding the relationship between structure and bioavailability in the literature.

Mechanistic based absorption models require in vitro information such as solubility and permeability and cannot be used for the purpose of early stage library design unless a quantitative model is developed for each model parameter (3). Lipinski's Rule of Five is the first qualitative attempt to develop tools to help chemists design bioavailable compounds (4). It established limits on *properties* such clogP, molecular weight, and number of hydrogen bond donors and acceptors, beyond which oral activity is predicted to be poor. Since its introduction, Lipinski's Rule of Five has been widely used in library design and candidate selection despite the fact that it produces a lot of false positive results. Similar, but more complex qualitative models have been recently reported in the literature to identify drug like molecules (5,6).

The aim of the present work is to develop a novel quantitative structure-bioavailability relationship (QSBR) to explore our ability to predict human bioavailability based on *structure*. Given a structure, the model will yield a bioavailability value. We will show that predictions made the QSBR model are more accurate than those made with Lipinski's Rule of Five.

METHODS

Bioavailability Database

Data for human oral bioavailability were obtained from the literature and an internal database (7,8). The generic drug names and the associated bioavailability values as well as experimental errors (if available) were entered into an electronic database so that a structure-bioavailability model could be created. SMILES strings were retrieved from the World Drug Index (WDI, Derwent Publishers, London) or created manually. Finally, 591 structures with SMILES, generic name, and bioavailability value were obtained. Any compounds whose bioavailability is strongly affected by the dose and formulation was excluded from the data set.

Quantitative Structure-Bioavailability Relationship (QSBR)

Development Procedure

SAS version 6.11 for IRIX 5.3 (SAS Institute Inc, Cary, NC) was used for model building as well as Splus version 3.4 Release 1 (MathSoft Inc, Seattle, WA). The statistical aim was to correlate bioavailability with molecular structure. Each SMILES structure was represented in terms of a "fingerprint" consisting of 608 substructure counts. Each substructure definition was defined using the SMARTS language from Daylight Chemical Information Systems, Inc (Santa Fe, NM). An in-house C program was used to pass each SMILES through the set of 608 substructure definitions to produce a string of substructure counts for each molecule. The counts are integer descriptors. The table of counts for all molecules (591×608) was then read into SAS statistical analysis software. Additional

[1] GlaxoWellcome Inc., Five Moore Drive, Research Triangle Park, North Carolina 27709.

[2] National Institute of Environmental Health Sciences, 111 Alexander Drive, MS D2-04, Research Triangle Park, North Carolina 27709.

[3] Present address: Food and Drug Administration, Division of Product Quality Research, 5600 Fishers Lane, HFD-941, NLRC 2400B, Rockville, Maryland 20857.

[4] To whom correspondence should be addressed. (e-mail: yul@cder.fda.gov)

0724-8741/00/0600-0639$18.00/0

variables were defined using recursive partitioning. The stepwise regression procedure (PROC REG) was used to construct a model for bioavailability based on the most significant fragment counts.

Recursive Partitioning

To improve the regression analysis, interactions between descriptors (9) were studied using recursive partitioning, a method that splits the bioavailability data into homogeneous groups (bins, partitions) in a hierarchical fashion and as a function of the descriptors to create a decision tree for the bioavailability. Whether or not a data split occurs is determined by the p-value. KnowledgeSEEKER version 4.1 (www.angoss.com) and Golden Helix Datamining (www.goldenhelix.com) software programs were used for recursive partitioning. For the purpose of this study, the initial two splits of the data (starting from the root bin) were used to find pairwise descriptor interactions that might impact the regression. The strategy was to find bins whose mean bioavailability was considerably different from the mean for the whole data set. The boolean classification rules *after two splits* for each bin define a pairwise descriptor interaction that might be used in a regression analysis.

Model Validation

The SAS software produces two kinds of predictions for the QSBR model, leave-everything-in versus leave-one-out (Press). In the case of leave-everything-in predictions, all compounds are considered in the model building and the resulting model predicts the bioavailability of each compound. In the case of leave-one-out, a model is built after removing one compound and the resulting model is used to predict the bioavailability of the one removed. This is repeated to obtain a prediction for every compound.

In addition to the leave-one-out validation, a more rigorous cross-validation was performed by randomly splitting the 591 observations into a training set (approximately 80% of the data) and a prediction set (approximately 20%). Model coefficients were then estimated using the training set, and this model was use to make predictions for compounds in the prediction set. The model root mean square error (RMSE) and the cross-validated R^2 for the prediction (20%) set were calculated for each round of validation, and 2000 such rounds were carried out.

Lipinski's Rule of Five Analysis

The Rule of Five predicts that oral activity is likely to be poor when there are more than 5 H-bond donors, 10 H-bond acceptors, the molecular weight is greater than 500, and/or the calculated Log P is greater than 5 (4). We sought to apply the Rule of Five to our experimental database to test its validity against the experimental bioavailabilities. However, the Rule of Five predictions are only qualitative, either good or poor. To compare these predictions with our experimental data, we must convert each experimental %F value into a *qualitative* good/poor value. A conservative criterion of 20% was used to classify the experimental %F values into good or poor values.

Comparison of QSBR with Lipinski's Rule of Five

We also sought to compare the QSBR model predictions against the qualitative Rule of Five predictions. Since the QSBR predictions are quantitative (predicts a %F value), they too must be converted to a good/poor value. Hence we again applied the 20% cutoff to differentiate between good and poor predictions. Leave-one-out predictions from the QSBR model were used for comparison. The use of leave-one-out made our predictions as realistic as possible and the comparison with the Rule of Five as impartial as possible.

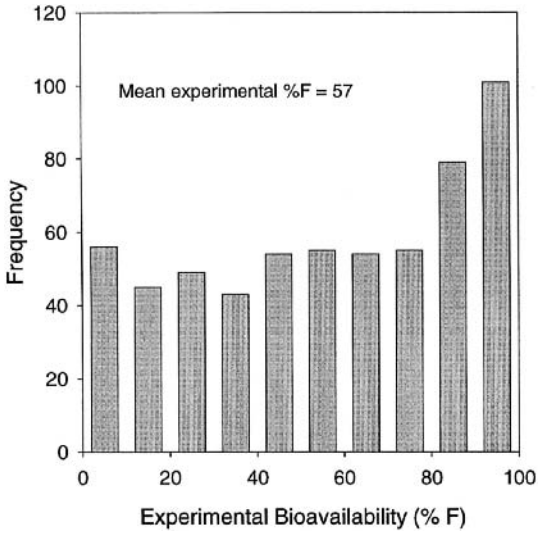

Fig. 1. The distribution of experimental human bioavailability data for 591 compounds.

RESULTS AND DISCUSSION

Descriptive Statistics of Human Oral Bioavailability

The histogram of the experimental bioavailability data (Fig. 1) showed that the data set was relatively well balanced with respect to high and low bioavailability. The mean experimental human bioavailability is 57. Experimental errors were obtained on 282 of the 591 compounds. The mean experimental error is 12 (Fig. 2), which is rather large and limits the accuracy

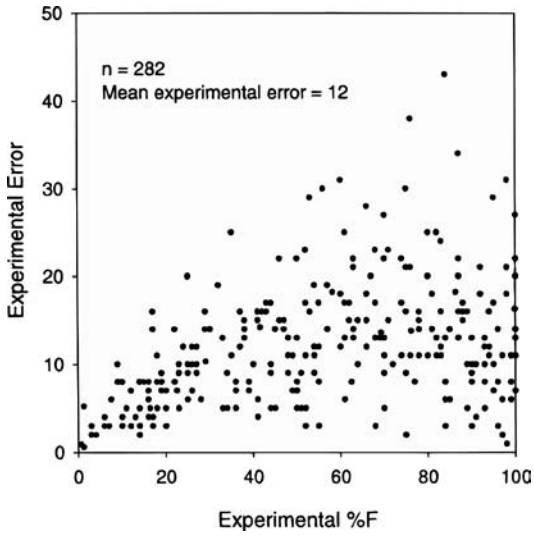

Fig. 2. Experimental error versus experimental bioavailability for 282 compounds.

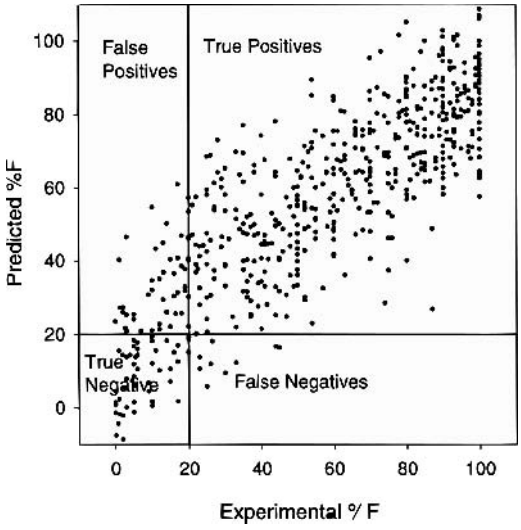

Fig. 3. Predicted versus experimental bioavailability from the QSBR model using leave-one-out predictions.

of any predictive models. Figure 2 shows that experimental error generally increases with increasing bioavailability.

Quantitative Structure-Bioavailability Relationship Model

The QSBR model obtained by stepwise regression had an R^2 of 0.71 and contained 85 substructural descriptors. The RMSE, an estimate of the error in the model, is 18. Given the mean experimental error of 12, this model error is reasonable. The cross-validated (PRESS, leave-one-out) R^2 is 0.63, indicating that unique compounds are not a particularly bad problem in the data set. A predicted versus actual plot is shown in Fig. 3 using leave-one-out predictions. The ratio of observations to descriptors is 591/85 or approximately 7, indicating that the model is not overfit.

The results for the full model and for the 80/20 cross-validation studies are summarized in Table I. The 80/20 results are again consistent with those for the full model. The mean cross-validated R^2 of 0.58 for the 80/20 splits is lower than the value of 0.63 obtained by leave-one-out validation, but this is to be expected since fewer observations are used to train the model. Further, the R^2 of 0.58 is not much below 0.6 that we regard as adequate proof of predictability.

The interactions between descriptors were studied using recursive partitioning. In principle, there are $(608)^2$ possible pairwise descriptor interactions given 608 descriptors. Use of recursive partitioning allowed us to find three interactions significant in the regression model. Figure 4 shows one of these interactions. At the top, the whole bioavailability database is contained in one bin. This represents the "root" of the tree. Recursive partitioning then finds significant ways to split the bioavailability data, considering all 608 descriptors. At the first split, X121 splits the bioavailability data set into five branches. X121 is a descriptor representing the number of carbons, a measure of how large the molecule is. Branch #1 contains those molecules that have between 1 and 6 carbon atoms ($1 <= X121 < 6$). Branch #2 contains those molecules that have between 6 and 16 carbons ($6 <= X121 < 16$), and etc. Branch #1 is then split by X278 into two branches. X278 represents the number of hydrogen bond acceptors. Branch #1 contains those molecules that have between 0 and 1 hydrogen bond acceptor atoms ($X278 < 1$). Branch #2 contains those molecules that have more than 1 acceptor atom ($X278 > 1$).

Notice that when the X121 split creates the first Branch #1, it creates a group of 18 molecules that have an average bioavailability of 45.91. This is not too different from the average value of the entire database, 57. However, when the X278 split creates the second Branch#2, it creates a smaller group of 10 molecules that have an average bioavailability of 18.60. This value is significantly lower than the database average of 57. One can conclude that the two splits that created this group of 10 molecules have created a "rule" that defines low bioavailability. Furthermore, the "interaction" under examination is between descriptors X121 and X278. The rules for each branch "interact" (or combine in a boolean AND sense) to create a more specific rule that represents a definition of low bioavailability. The specific rule in Fig. 4 is that if #carbons <6 and #H-bond acceptors >1, then bioavailability will be lower than average. If this rule is added to the regression problem in the form of an interaction descriptor (shown in entry 34, Table III), the regression procedure detects that the bioavailability is lower than average when that descriptor is present and assigns a negative regression coefficient to it.

Table I. Results for Quantitative Structure-Bioavailability Model

Name	Results
Number of descriptors	85
Model R^2	0.71
Root mean squared error (RMSE)	17.92
Cross validated (leave-one-out) R^2	0.63
Mean Cross validated (80/20 splits) R^2 (20% sets)	0.58*
Mean RMSE for prediction (20% sets)	20.40*

* Averaged over 2000 splits.

Lipinski's Rule of Five Predictions

Using our compilation of 591 experimental %F values, we found that 490 compounds have good bioavailability while 101 compounds have poor bioavailability. The Rule of Five correctly predicts 462 of 490 good bioavailability compounds and 22 of 101 poor bioavailability compounds. Five out of the total 591 compounds can not be computed for the Rule of Five properties and thus can not be predicted by the Rule of Five.

Comparison of QSBR with Lipinski's Rule of Five

It is important to be able to filter out from the drug screening process compounds that are non-bioavailable. Of particular interest are false positive predictions, meaning compounds that are predicted to be bioavailable but that are experimentally non-bioavailable. Compounds in this category should not be developed but are predicted to be bioavailable; they would be developed and would probably fail, thus driving up the cost of drug development unnecessarily.

Also of particular interest are the false negatives, compounds that should be developed but that are predicted to be

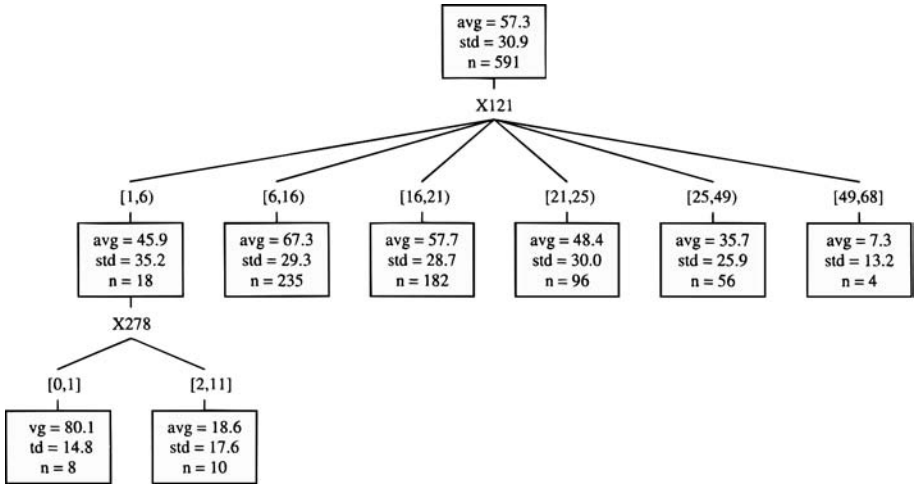

Fig. 4. The decision tree that defines the boolean interaction of desciptors, X121 and X278 (entry 34 in Table III).

not bioavailable and are therefore discarded. False negatives would be serious problems in both the discovery and development settings. In the first case, you eliminate a possible drug lead, while in the second case you eliminate a possible drug that has demonstrated biological potency.

Table II shows the predictions of the human oral bioavailability by the QSBR model and Lipinski's Rule of Five. Lipinski's model predicts 95% of the positives (5% false negatives) correctly but only 22% of the negatives correctly (78% false positives). The QSBR model predicts 97% of the positives correctly (3% false negatives) and 47% of the negatives correctly (53% false positives). This represents an improvement in the prediction of bioavailability.

Descriptors in QSBR Model

The descriptors used are either substructure counts (base 10 integers) or combinations of them. Table III is a partial list of descriptors used in the QSBR model and includes the regression coefficient. The magnitude of each coefficient is a measure of its relative impact upon bioavailability, and its sign indicates whether bioavailability generally increases or decreases as that particular fragment count increases. Improvement of molecular bioavailability is dependent upon increasing the use of fragments with positive coefficients.

The descriptor with the lowest p-value is the number of heavy atoms (non-hydrogens) in the molecule. This descriptor has a small negative coefficient but is important due to the large number of heavy atoms in a typical drug. Because of the negative coefficient, small molecules should be more bioavailable than large ones, in agreement with Lipinski's Rule of Five molecular weight cutoff. Each heavy atom reduces the %F value by about one percentage point. The hydrogen bond donor descriptor carries a negative coefficient (decreased bioavailability) but the hydrogen bond acceptor descriptor carries a positive coefficient (increased bioavailability).

The worst fragments for bioavailability are tetrazole, 4-aminopyridine, and benzoquinone. Other detrimental fragments are dihydropyran and cyclohexanone. Some of the best fragments for bioavailability are azide, 1-methylcyclopentyl alcohol, salicylic acid, and cyanoguanidine. The halogens seem to have small positive coefficients. An N-terminal amino acid residue has a positive coefficient whereas an interior amino

Table II. Predictions of Human Oral Bioavailability by the QSBR model and Lipinski's Rule of Five. Percentages Are Column Percentages*

Experimental Bioavailability	Number of Compounds	QSBR Model Prediction		Lipinski's Rule of Five**	
		Good	Poor	Good	Poor
Good (%F > 20)	490	476 (97%) True Positive	14 (3%) False Negative	462 (95%) True Positive	25 (5%) False Negative
Poor (%F < 20)	101	54 (53%) False Positive	47 (47%) True Negative	77 (78%) False Positive	22 (22%) True Negative

* True Positive: Experimental good bioavailability, predicted good bioavailability, too. False Positive: Experimental poor bioavailability, but predicted good bioavailability. True Negative: Experimental poor bioavailability, predicted poor bioavailability, too. False Negative: Experimental good bioavailability, but predicted poor bioavailability.

** 3 out of 490 good bioavailability compounds and 2 out of 101 poor bioavailability compounds cannot be computed for the Rule of Five properties.

Table III. Partial List of Descriptors Used in the QSBR Model, Sorted by Regression Coefficient

No	Name	SMARTS language definition	Regression coefficient
1	tetrazole	[nH]1nnnc1	−73
2	4-aminopyridine	[nX2]1c([CX4,c,H])c([CX4,c,H])c(N)c([CX4,c,H])c1([CX4,c,H])	−62
3	benzoquinone	O=[C,c]1[C,O,c]~[C,c][C,c](=O)[C,c]~[C,c]1	−55
4	dihydropyran	O1CCCcc1	−40
5	quaternized pyridinium	[CX4,c][n&+]1acccc1	−36
6	cyclohexanone	O=C1[C,O]CCC~C1	−31
7	sulfhydryl group	[SX2;H1][#6]	−23
8	thioether	[SX2](−[A,a])−[A,a]	−21
9	divalent nitrogen	[NX2;!R]	−20
10	primary amide	[CX3](−[C,c])(−[NH2])=O	−18
11	tertiary amide	[N](−[CX4])(−[CX4])−C=O	−17
12	tertiary amine	[NX3;H0](−a)(−a)−[A,a]	−13
13	tertiary amine	[NH0]([CX4])([CX4])[CX4]	−13
14	aromatic, aliphatic ketone	[CX3](−c)(−C)=O	−12
15	interior amino acid residue	[O]=[C,S,P]−*−[NH]	−12
16	tertiary amine	[NH0]([CH3])([CH3])[CX4]	−6.5
17	hydrogen bond donor	H−[N,O,S]	−6.3
18	any heavy atom	[A,a]	−0.8
19	fluorine	F	2.27
20	hydrogen bond acceptor	[$([NX3;H0](−[CX4])(−[CX4])[CX4]),$([nX2](:c):c), $(O=[C,S,P]),$([OX2](−[CX4])−[CX4]),$([O&−])]	4.5
21	iodine	I	8.18
22	N-terminal amino acid residue	[O]=[C,S,P]−*−[NH2]	10.7
23	any amide	[NX3]C(=O)[#6]	13.1
24	alkanoic acid	[CH2][CH2]−C(=O)[OH]	13.8
25	thioether	[SX2]([#6])[#6]	14.5
26	cyclopropyl	[CH]1[CH2][CH2]1	16.1
27	aromatic, aliphatic ester	[CX3](−c)(−O−C)=O	18.3
28	cyanoguanidine	[C](−N)(−N)=N−C#N	27.8
29	salicylic acid	c1([OH])ccccc1C(=O)[OH]	29.9
30	1-methylcyclopentyl alcohol	C1CC~CC1([OH])[#6]	47.9
31	azide	N=[N+]=[N−]	56.7
32	acids	if #COOH >1 or if #strong acids >0, then poor	−24
33	small and polar molecule	if #carbons <=16 and #hydroxyls >2, then poor	−30
34	small and polar molecule	if #carbons <6 and #H-bond acceptors >1, then poor	−35

acid residue has a negative coefficient. An aromatic, aliphatic ketone is less bioavailable than an aromatic, aliphatic ester.

Primary and tertiary amides seem to have a negative coefficient, but these coefficients are mostly balanced by a positive coefficient for amides in general. This example demonstrates that linear combinations of related definitions are effective as regressors. Other examples of this exist in the descriptor set. Thioethers are represented by two descriptors, the sum of the coefficients being somewhat negative. Another example involves azide and divalent nitrogen.

Almost 20% of the 591 compounds in the database contain a carboxylic acid and most of these compounds have high bioavailability. However, only one type of carboxylic acid appears in the list of descriptors, an alkanoic acid fragment that has a positive coefficient. That carboxylic acids do not have a more general representation in the regression model is a reflection of the fact that one can attain high bioavailability without a carboxylic acid. Nonetheless, an indicator variable was constructed that proved to be highly significant in the model. If *more than one* carboxylic acid appears, or if a strongly acidic group (sulfonic or phosphoric) appears at all, then bioavailability should suffer.

CONCLUSIONS

A novel quantitative structure-bioavailability relationship has been developed to predict human oral bioavailability based on molecular structure. As compared to Lipinski's Rule of Five, the QSBR model gives a lower percentage of false positive predictions. The substructural descriptors resulted from the work can be used to guide chemists on how to increase oral bioavailability in humans.

ACKNOWLEDGMENTS

We would like to thank Darko Butina and Gianpaolo Bravi for assistance with computation of fingerprints, Elaine Hopkins and Barbara Reitter for assistance with the compilation of human bioavailability data, and David Cummins and Stan Young for discussions concerning recursive partitioning.

REFERENCES

1. R. A. Fecik, K. E. Frank, E. J. Gentry, S. R. Menon, L. A. Mitscher, and H. Telikepalli. The search for orally active medications through combinatorial chemistry. *Med. Res. Rev.* **18**:149–185 (1998).
2. PriceWaterHouseCoopers. Pharma 2005: An Industrial Revolution in R&D. 1999.
3. L. X. Yu. An integrated absorption model for determining dissolution, permeability, and solubility limited absorption. *Pharm. Res.* **16**:1884–1888 (1999).
4. C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev.* **23**:3–25 (1997).
5. Ajay, W. P. Walters, and M A. Murcko. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **41**:3314–3324 (1998).
6. J. Sadowski and H. Kubinyi. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**:3325–3329 (1998).
7. W. K. Sietsema. The absolute oral bioavailability of selected drugs. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **27**:179–211 (1989).
8. L. Z. Benet, S. Øie, and J. B. Schwartz. Design and optimization of dosage regiments; pharmacokinetic data. In J. G. Hardman, L. E. Limbird, A. G. Gilman (eds). *Pharmacological Basis of Therapeutics*, 9th ed., McGraw-Hill: New York, 1996, pp. 1707–1793.
9. J. M. Chambers and T. J. Hastie. Statistical Models. In S. J. M. Chambers and T. J. Hastie (eds.). *Statistical Models*, Wadsworth & Brooks/Cole Advanced Books and Software, Pacific Grove, 1992, p. 22.