In Silico Renal Clearance Model Using Classical Volsurf Approach

Munikumar R. Doddareddy, Yong Seo Cho, Hun Yeong Koh, Dong Hyun Kim, and Ae Nim Pae*

Life Science Division, Korea Institute of Science and Technology, P.O. Box 131,
Cheongryang, Seoul 130-650, Korea

Received August 16, 2005

A data set of 130 diverse compounds containing both central nervous system (CNS) and non-CNS drugs was used to generate a renal clearance model using a classical Volsurf approach. Percentage renal clearance data was used as a biological input. The score plots obtained from principal component analysis and partial least-squares (PLS) analysis clearly separated high-clearance compounds from low-clearance compounds. PLS models were used to predict the renal clearance of the data set. Categorical statistical methods such as SIMCA and recursive partitioning techniques were used for classifying the compounds into low- and high-clearance categories. PLS coefficient plots, Volsurf descriptor profiles, 3D Grid maps, and RP decision trees were used to explain the important descriptors separating low and high renal clearance compounds. For comparative purposes, topological descriptors such as Molconn-Z were also examined. All the models were validated by an external test set of 20 compounds. These models can be used as efficient tools in the classification and prediction of the renal clearance of unknown compounds, the knowledge of which is helpful in understanding their bioavailability behavior.

INTRODUCTION

The high-speed generation of compounds by using modern drug discovery tools such as combinatorial chemistry and recent advances in lead compound identification using high-throughput and in silico techniques has allowed the rapid identification of compounds exhibiting possible pharmacological effects at known drug receptor sites. However, successful drug candidates must also possess other attributes to make them suitable for clinical application. These drugs must be ultimately administered to humans, and properties such as human toxicity, bioavailability, and other pharmacokinetic parameters become very important. In vitro tests using animal models for determining absorption, distribution, metabolism, and excretion properties and toxicity are often time-consuming and expensive.[1] Even then, the results may not accurately reflect human pharmacokinetics, and it has been reported that the majority of drugs dropped from development were dropped because of efficacy or pharmacokinetic difficulties.[2] There is need for more progress in the generation of in silico models useful for the prediction of human pharmacokinetic parameters, which can directly enhance the success of drug development programs. In silico methods to predict the pharmacokinetic properties of drugs have been recently reviewed by Ekins et al.[3] The ultimate aim of many of these studies is to be able to predict the bioavailability of drugs.

Methods of generating descriptors solely from drug structure are gaining popularity because of their resource-saving potential and success in quantitative structure−activity and structure−property relationship (QSAR and QSPR) analyses. These descriptors range in complexity from simple one-dimensional atomic and functional group counts to two-dimensional topological and charge indices to complex three-dimensional descriptors which often rely on conformational aspects of a molecule. Both one- and two-dimensional topological indices have been used extensively to numerically relate molecular structure to activity and properties.[4] These descriptors rely only on the molecular graph for their calculation. In contrast, three-dimensional descriptors require the absolute conformation of a molecule to be described, and the information gained is specific to that conformation. They also have been successfully used to develop QSPRs.[5] Various methods for constructing QSAR/QSPR models have been used including multilinear regression,[6] principal component analysis (PCA),[7] and partial least-squares (PLS)[8] regression. In addition, artificial neural networks have become popular due to their success where complex nonlinear relationships exist among data, as is often the case when dealing with drug data sets.[9]

Yoshida and Topliss[10] and Andrews et al.[11] recently presented successful QSAR models for human bioavailability data, based on a large number of structurally diverse drugs (232 and 591, respectively). Despite the apparent success of the models, it is obvious that the accurate prediction of bioavailability is an ambitious ultimate aim. A more successful approach may be the prediction of the components of bioavailability. These include absorption, distribution, tissue binding, metabolism, and excretion; they are usually parametrized individually using in vitro models. Attempts to improve the bioavailability profile of a drug usually focus on improving one of these processes. One of them is to reduce metabolic clearance,[12] the bulk of which takes place in the liver. In silico models obtained from the human clearance data can deliver models that correspond more to the in vivo situation where a compound is subject to a variety of metabolizing enzymes. The main purpose of this study was to build a model to predict drug biotransformation in

* Corresponding author. Tel.: +82-2-958-5185. Fax: +82-2-958-5189. E-mail: anpae@kist.re.kr.

10.1021/ci0503309 CCC: $33.50 © 2006 American Chemical Society
Published on Web 02/10/2006

humans as a multienzymatic, composite process. The assumption made in this study was that the excretion of the unchanged drug takes place only by a urinary route. Urinary clearance (the amount of drug excreted unchanged in the urine) gives a good means to monitor the proportion of a drug that is subject to metabolism. To achieve this aim, the quantity of drug excreted unchanged in the urine in humans, expressed as a percent of the administered dose, was taken as the endpoint representing no biotransformation. The present study is dedicated to deriving chemometric models for a data set of structurally diverse drugs via a classical Volsurf approach[13] to understand the effect of Volsurf descriptors[14] on the renal clearance of these compounds. For a comparative study, we have also studied topological descriptors such as Molconn-Z[15] and categorical statistical analysis methods such as the SIMCA[16] (soft independent modeling by class analog) and RP[17] (recursive partitioning) techniques. The generated models were intended to be used for predicting the renal clearance of our library compounds.

Molecular descriptors calculated by the Volsurf program have been extensively used to model pharmacokinetic properties, for example, passive permeability through the gastrointestinal tract or through the blood−brain barrier.[18] These descriptors quantify steric, hydrophobic, and hydrogen-bond interactions between model compounds and different environments. Volsurf is an automatic procedure for the conversion of 3D molecular fields into physicochemically relevant molecular descriptors. In the standard procedure, the interaction fields with a water probe and a hydrophobic probe are calculated for all molecules in the data set. However, grid maps produced by other probes (ionic probes etc.) or by various molecular mechanics or semiempirical approaches (e.g., electrostatic potential) can also be used. The basic concept of Volsurf is to extract the information present in 3D molecular field maps into a few quantitative numerical descriptors, which are easy to understand and interpret. Molecular recognition is achieved using image analysis software coupled with external chemical knowledge. Within this context, Volsurf selects the most appropriate descriptors and parametrization according to the type of 3D maps under study.

The molecular descriptors obtained refer to molecular size and shape, to the size and shape of both hydrophilic and hydrophobic regions, and to the balance between them. Hydrogen bonding, amphiphilic moments, and critical packing parameters are other useful descriptors. The Volsurf descriptors have been presented and explained in detail.[13] The originality of Volsurf resides in the fact that surfaces, volumes, and other related descriptors can be directly obtained from 3D molecular fields with simple computational algorithms. Moreover, Volsurf descriptors can be easily obtained for small, medium, and large molecules, as well as for biopolymers such as DNA sequences, peptides, and proteins. No parametrization is required. Furthermore, it was already demonstrated that the Volsurf descriptors are often hardly influenced by conformational sampling and averaging.[13,14] This is probably due to the peculiarity of the Grid force field that allows for the conformational flexibility of external groups, hydrogens, and lone pairs. In general, a protocol consisting of a simple 2D-to-3D structure conversion followed by energy minimization produces good results, without any molecular dynamics sampling and Boltzmann averaging. These properties make Volsurf descriptors computationally efficient and well-suited for fast quantitative structure−property relationship studies, especially when dealing with a large number of compounds.

Molconn-Z's families of descriptors[15] are based on the pioneering work of Lowell Hall and Lemont Kier.[19] These descriptors reflect the intrinsic features of the molecule atoms, their electronic state, and their relation to one another. Molconn-Z descriptors can be readily used with standard statistical methods to create QSAR and QSPR models that predict the biological activity or physical properties of new molecules.

MATERIALS AND METHODS

Data Set. A data set of 150 diverse compounds, 130 of which were used as a training set and 20 of which were used as a test set, were obtained from ref 20. The names of the compounds used for the analysis are given in Table 1 along with urinary excretion data. Care was taken to include compounds such as cefadroxil and cephalexin, which have high percentages of renal clearance of more than 90%, and also compounds such as doxipen or prazepam, which have zero renal clearance. 2D to 3D conversion was carried out by the program Concord,[21] and the resulting structures were given Gasteiger−Marsili charges. Energy minimization was done using a Tripos force field[22] with a distance-dependent dielectric and a conjugate gradient method with a convergence criterion of 0.01 kcal/mol as implemented in SYBYL 7.0.[23]

CALCULATION OF DESCRIPTORS

A set of 94 Volsurf descriptors[14] was automatically generated from 3D molecular fields by using the Volsurf 4.1 program.[24] The water probe (OH2) was used to simulate solvation−desolvation processes, while the hydrophobic probe (DRY) and the carbonyl probe (O) were used to simulate drug−membrane interactions. The DRY probe is a specific probe for the computation of the hydrophobic energy; the overall energy of the hydrophobic probe, R, is computed at each gridpoint as $E_{\mathrm{entropy}} + E_{\mathrm{LJ}} - E_{\mathrm{HB}}$, where E_{entropy} is the ideal entropic component of the hydrophobic effect in an aqueous environment, E_{LJ} measures the induction and dispersion interactions occurring between any pair of molecules, and E_{HB} measures the H-bonding interactions between water molecules and polar groups on the target surface.

A set of 95 Molconn-Z descriptors[15] was generated by using the Molconn-Z module given in the SYBYL program.[23] Molconn-Z descriptors mainly include molecular connectivity (χ) indices; shape (κ) indices; electrotopological state (E-state) indices (comprised of atom-, bond-, and group-type descriptors); topological state (including Shannon and Wiener) and equivalence indices; and counts of graph paths, atoms, atoms types, bond types, rings, and so forth. To increase the predictive ability of the models, correlation matrixes of the descriptors were built and the number of descriptors was filtered to 62 Volsurf descriptors and 37 Molconn-Z descriptors on the basis of a correlation threshold of 0.9.

Statistical Analysis. PCA[7] and PLS[8] are the most common chemometric tools for extracting and rationalizing the information from any multivariate description of a biological

Table 1. Renal Clearance Data[20] Used for the Analysis

compound	exptl[a]	calcd[a,b]	resi[a,b]	calcd[a,c]	resi[a,c]	compound	exptl[a]	calcd[a,b]	resi[a,b]	calcd[a,c]	resi[a,c]
acebutolol	40.00	29.91	10.09	15.91	24.09	clemastine	2.00	2.52	−0.52	8.43	−6.43
alfentanil	1.00	2.96	−1.96	11.07	−10.07	clonazepam	1.00	11.12	−10.12	5.95	−4.95
alprazolam	20.00	9.37	10.63	9.03	10.98	clorazepate	1.00	10.79	−9.79	21.82	−20.82
amitriptyline	2.00	2.65	−0.65	−7.04	9.04	cyclophosphamide	6.50	9.10	−2.60	−4.89	11.39
amlodipine	10.00	13.19	−3.19	15.71	−5.71	dapsone	15.00	7.69	7.31	47.48	−32.48
amoxicillin	86.00	84.84	1.16	77.64	8.36	desipramine	2.00	0.91	1.09	4.40	−2.40
ampicillin	82.00	67.54	14.46	73.65	8.36	desmethyldiazepam	1.00	−7.72	8.72	3.10	−2.10
aztreonam	68.00	75.93	−7.93	99.62	−31.62	dexamethasone	2.60	19.93	−17.33	12.20	−9.60
bepridil	1.00	−6.23	7.23	12.01	−11.01	diazepam	1.00	−5.18	6.18	−4.43	5.43
betaxolol	15.00	12.91	2.09	21.89	−6.89	diclofenac	1.00	7.61	−6.61	19.30	−18.30
bumetanide	62.00	70.92	−8.92	53.80	8.20	dicloxacillin	60.00	45.20	14.80	51.26	8.75
bupropion	1.00	11.33	−10.33	8.35	−7.35	didanosine	36.00	55.79	−19.79	36.42	−0.42
busulfan	1.00	−0.75	1.75	8.80	−7.80	diphenhydramine	1.90	2.30	−0.40	1.33	0.57
caffeine	1.10	14.04	−12.94	17.51	−16.41	doxepin	0	5.83	−5.83	−8.04	8.04
captopril	38.00	26.98	11.02	32.47	5.53	enoxacin	45.00	34.03	10.97	31.18	13.82
carbenicillin	77.00	54.45	22.55	59.30	17.70	esmolol	1.00	5.48	−4.48	8.44	−7.44
cefaclor	52.00	72.10	−20.10	60.91	−8.91	felodipine	1.00	3.17	−2.17	15.18	−14.18
cefadroxil	93.00	102.8	−9.87	61.40	31.60	flecainide	43.00	50.54	−7.54	28.20	14.80
cefazolin	80.00	68.91	11.09	61.49	18.51	fluconazole	75.00	51.35	23.65	52.62	22.38
cefotaxime	55.00	44.59	10.41	78.40	−23.40	fluoxetine	2.50	15.68	−13.18	9.89	−7.39
cephalexin	91.00	80.90	10.10	57.25	33.75	flurbiprofen	2.00	15.34	−13.34	8.46	−6.46
cephalotin	52.00	33.75	18.25	49.54	2.47	furosemide	66.00	61.87	4.13	68.26	−2.26
cephapirin"	48.00	45.69	2.31	32.66	15.34	ganciclovir	73.00	87.63	−14.63	64.33	8.67
chlorothiazide	92.00	77.44	14.56	54.39	37.61	gemfibrozil	1.00	−1.49	2.49	8.72	−7.72
chlorpheniramine	13.00	7.34	5.66	2.35	10.65	glyburide	0.10	10.71	−10.61	23.72	−23.62
chlorpropamide	20.00	25.84	−5.84	31.92	−11.92	haloperidol	1.00	2.52	−1.52	4.90	−3.90
chlortalidone	65.00	45.31	19.69	55.62	9.38	hexobarbital	1.00	5.41	−4.41	2.89	−1.89
cinoxacin	72.50	50.91	21.59	26.60	45.90	imipramine	2.00	8.03	−6.03	2.54	−0.54
ciprofloxacin	65.00	35.81	29.19	43.24	21.76	isotretinoin	1.00	5.40	−4.40	7.50	−6.50
clavulanicacid	43.00	54.19	−11.19	43.79	−0.79	ketamine	4.00	3.03	0.97	2.81	1.19
ketoprofen	1.00	−9.06	10.06	8.90	−7.90	oxaprozin	0.10	1.46	−1.36	7.80	−7.70
ketorolac	7.50	5.07	2.43	26.40	−18.90	oxazepam	1.00	13.93	−12.93	16.91	−15.91
lidocaine	2.00	2.80	−0.80	7.23	−5.23	paroxetine	2.00	−0.26	2.26	−10.1	12.12
loratadine	0.10	−5.05	5.15	−27.6	27.70	pentoxifylline	0	19.35	−19.35	13.62	−13.62
lorazepam	1.00	11.34	−10.34	17.80	−16.80	phenobarbital	24.00	19.22	4.78	7.25	16.75
lorcainide	2.00	−8.38	10.38	−7.95	9.95	phenylbutazone	1.00	−7.30	8.30	1.82	−0.82
melphalan	12.00	10.94	1.06	20.19	−8.19	phenytoin	2.00	−7.00	9.00	11.64	−9.64
meperidine	12.50	−4.64	17.14	−4.77	17.27	pindolol	54.00	27.67	26.33	28.22	25.78
methotrexate	81.00	78.51	2.49	82.89	−1.89	prazepam	0	−7.78	7.78	8.77	−8.77
methylprednisolone	4.90	12.79	−7.89	13.15	−8.25	prednisolone	26.00	20.40	5.60	14.80	11.21
metoclopramide	20.00	12.09	7.91	32.52	−12.52	prednisone	3.00	18.57	−15.57	13.41	−10.41
metoprolol	10.00	16.75	−6.75	8.90	1.10	proguanil	50.00	35.52	14.48	58.32	−8.32
metronidazole	10.00	25.55	−15.55	27.13	−17.13	propafenone	1.00	18.66	−17.66	4.27	−3.27
mexiletine	9.50	9.12	0.38	23.28	−13.78	propylthiouracil	2.00	8.82	−6.82	20.77	−18.77
morphine	4.00	12.39	−8.39	11.65	−7.65	quinidine	18.00	9.94	8.06	2.96	15.04
nafcillin	43.00	29.52	13.48	40.31	2.69	risperidone	3.00	10.89	−7.89	−8.62	11.62
nalbufine	4.00	9.45	−5.45	29.39	−25.39	salicylicacid	16.00	36.24	−20.24	21.92	−5.92
naloxone	0.10	−5.08	5.18	7.12	−7.02	simvastatinacid	0.10	−6.06	6.16	5.92	−5.82
naltrexone	1.00	−8.39	9.39	17.64	−16.64	sotalol	75.00	72.79	2.21	31.62	43.38
naproxen	1.00	9.49	−8.49	6.84	−5.84	sufentanil	6.00	−5.30	11.30	3.79	2.21
nicardipine	1.00	11.69	−10.69	−4.74	5.74	sulfisoxazole	49.00	43.77	5.23	43.83	5.17
nicotine	16.70	9.90	6.80	22.88	−6.18	sulpiride	70.00	70.22	−0.22	55.68	14.33
nifedipine	0	8.20	−8.20	1.12	−1.12	sumatriptan	22.00	24.17	−2.17	25.57	−3.57
nimodipine	1.00	12.87	−11.87	−7.36	8.36	tamoxifen	1.00	−4.55	5.55	−18.9	19.98
nitrazepam	1.00	16.01	−15.01	5.01	−4.01	temazepam	1.00	6.65	−5.65	12.73	−11.73
nitrendipine	1.00	13.87	−12.87	−1.72	2.72	terazosin	12.00	21.48	−9.48	26.41	−14.41
nitrofurantoin	47.00	40.86	6.14	27.44	19.56	terbutaline	56.00	48.90	7.10	30.32	25.68
nortriptyline	2.00	−0.22	2.22	−4.96	6.96	tetracycline	58.00	65.52	−7.52	70.39	−12.39
ondansetron	5.00	−0.88	5.88	0.94	4.06	thc	1.00	5.47	−4.47	−9.42	10.42
oxacillin	46.00	50.24	−4.24	47.77	−1.77	theophylline	18.00	40.66	−22.66	25.91	−7.91
thiopental	1.00	3.34	−2.34	20.76	−19.76	triazolam	2.00	10.04	−8.04	9.71	−7.71
timolol	15.00	37.13	−22.13	32.23	−17.23	venlafaxine	4.60	−3.92	8.52	3.12	1.48
tocainide	38.00	25.68	12.32	30.75	7.25	warfarin	2.00	−9.34	11.34	4.23	−2.23
tolbutamide	0.00	22.73	−22.73	28.45	−28.45	zalcitabine	65.00	45.62	19.38	53.64	11.36
trazodone	1.00	0.24	0.76	21.43	−20.43	zolpidem	1.00	3.78	−2.78	−7.42	8.42
						Test Set					
acetylsalicylicacid	1.40	20.85	−19.45	19.94	−18.54	flumazenil	1.00	17.15	−16.15	10.59	−9.59
acyclovir	75.00	82.88	−7.88	60.83	14.17	hydrochlorothiazide	95.00	88.33	6.67	56.80	38.20
atropine	25.00	14.53	10.47	14.06	10.94	indomethacin	15.00	−10.81	25.81	5.80	9.20
budesonide	0	3.82	−3.82	−5.66	5.66	isradipine	0	22.19	−22.19	6.72	−6.72
carbamazepine	1.00	18.27	−17.27	23.86	−22.86	lomefloxacin	65.00	41.97	23.03	31.76	33.24
cefamandole	96.00	73.28	22.72	50.33	45.67	mercaptopurine	22.00	42.20	−20.20	35.68	−13.68
coccaine	2.00	21.48	−19.48	7.37	−5.37	piroxicam	5.00	28.33	−23.33	29.13	−24.13
diazoxide	35.00	20.62	14.38	25.21	9.79	sertraline	1.00	19.16	−18.16	1.37	−0.37
doxycycline	41.00	59.22	−18.22	69.95	−28.95	ticarcillin	77.00	63.81	13.19	77.94	−0.94
ethambutol	79.00	87.14	−8.14	25.18	53.82	verapamil	3.00	−11.65	14.65	−22.78	25.78

[a] exptl, experimental; calcd, calculated; resi, residual. [b] Volsurf predictions. [c] Molconn-Z predictions.

system. Complexity reduction and data simplification are two of the most important features of such tools. PCA and PLS condense the overall information into two smaller matrixes, namely, the score plot and the loading plot.[25] The score plots represent the relative position of the objects in the space (two-dimensional or three-dimensional) of the principal components. They are useful to identify clusters of objects and single objects that behave in a peculiar way. Moreover, the position of the objects in the plots may serve to interpret the principal components (PCs). The first PCs try to explain the maximum amount of variation and, therefore, when there are clusters of objects, to distinguish among them. In this context, the PC can be interpreted as a compendium of distinctive features of the objects in these clusters. The loading plots represent the original variables in the space (two-dimensional or three-dimensional) of the principal components. The loading of a single variable indicates how much this variable participates in defining the PC (the squares of the loadings indicate their percentage in the PC). Variables contributing very little to the PCs have small loading values and are plotted around the center of the plot. On the other hand, the variables that contribute most are plotted around the borders of the plot. Because the chemical interpretation of score and loading plots is simple and straightforward, PCA and PLS are usually preferred to other nonlinear methods, especially when the noise is relatively high.

Score and loading plots are interconnected so that any descriptor change in the loading plot is reflected by changes in the position of compounds in the score plot. A pairwise comparison can be made directly with interactive plots,[26] as developed in the Volsurf program,[24] and the relative contributions to the property are shown in the related descriptors space.

PCA is a least-squares method, and for this reason, its results depend on data scaling. The initial variance of a column variable partly determines its importance in the model. To avoid this problem, column variables were scaled to unit variance before analysis.[27] The column average was then subtracted from each variable. From a statistical point of view, this corresponds to moving the multivariate system to the center of data, which becomes the starting point of the mathematical analysis. The same autoscaling and centering procedures were applied to the PLS analysis.

Once the PCA model was developed, PCA predictions for new compounds or external test set compounds were made by projecting the compound descriptors into the PCA model. This was accomplished by calculating the score vector **T** of descriptors X and average x for the new compounds, using the loading P of the PCA model, according to the following equation (eq 1):

$$\mathbf{T} = (X - x)P'(PP')^{-1} \tag{1}$$

For the PLS discrimination, external predictions were made using the following equation (eq 2):

$$Y = y - xP'(PP')B\mathbf{Q} + XP'(\mathrm{PP}')^{-1}B\mathbf{Q} \tag{2}$$

Where y is the Y column average, **Q** is the loading vector for the y space, and B is the coefficient between the X and Y spaces.

SIMCA[16] calculations are done using the SYBYL program.[23] SIMCA is a technique for developing a model which relates to known categories in a table on the basis of calculating a separate principal component for each category. It produces a mathematical description of the differences between rows of different categories, on the basis of columns of explanatory properties. SIMCA is conceptually the tool to use in place of PLS when the target property is categorical rather than a continuous variable. Cross validation is always performed internally, and the final model is returned for examination.

Table 2. Summary of PCA Analysis

components	XVarExp[a]	XAccum[b]
1	31.92	31.92
2	12.35	44.27
3	9.47	53.74
4	7.09	60.83
5	3.62	64.45

[a] XVarExp, percentage of X-matrix variance explained by that component. [b] XAccum, accumulative percentage of the X-matrix variance explained by the model

RP[17] (recursive partitioning) calculations are done using the QSAR module in the Cerius2[28] program. RP is a technique which can be applied to mine large data sets in order to uncover hidden patterns within the data and to elucidate statistically significant subgroupings within the data. In general terms, RP is a data-analysis method for relating a "dependent" variable (Y) to a collection of independent variables (X) in order to uncover or simply understand the elusive relationship, $Y = f(X)$. The central result of recursive partitioning is a "decision tree" or "graph", in which the data is organized (partitioned) into nodes along branches; data which are more similar according to some criteria tend to be localized into the same nodes, whereas more dissimilar data tend to occupy different nodes. The statistical significance of the "split" of the data into the nodes can be placed on a more quantitative footing by computing p values, which discern the quality of a split relative to a random event, and provides the value-added component to the data. In our work, the twoing rule was used to decide where to split in a growing tree. The twoing rule tends to split into two nodes with roughly the same number of examples, so the trees look more balanced, although the classification rates are sometimes slightly less. Recursive partitioning does not try to stop splitting at the right moment; instead, it is designed to "oversplit" and then prune the tree backwards. To control the amount of pruning, a moderate pruning option was selected. The minimum number of samples at each node was set to 5, and the number of cross-validation groups was set to 10.

RESULTS AND DISCUSSION

In this study, first, a PCA was conducted to search for the relationship between the 3D structure and the percent urinary excretion data of the training set containing 130 diverse sets of compounds (Table 1) and Volsurf descriptors. When a correlation matrix was used, 94 descriptors were filtered to 62 using a correlation threshold of 0.90. No biological input was given to the model. The first five components obtained by the cross-validation technique explained about 64.45% of the total variance of the matrix (Table 2). The score plot for the first two PCs is shown in

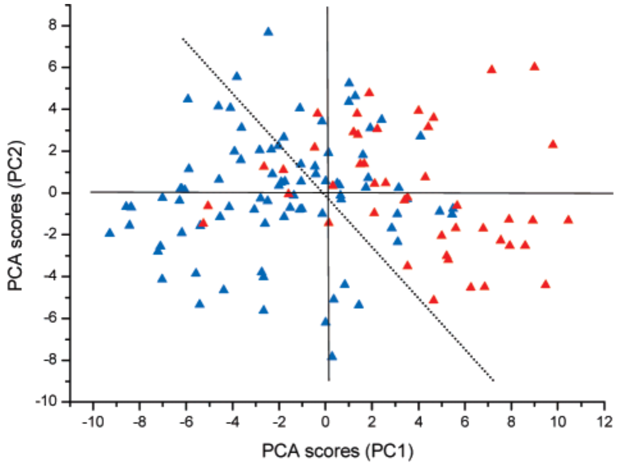

Figure 1. Score plot of PC1 vs PC2 for the PCA analysis. Training set compounds are colored according to percent renal clearance. Blue (0−20%) and red (20−100%).

Figure 1. To understand the plots clearly, the data set was divided into two groups and colored according to the percent renal clearance. A total of 85 compounds that have a percent renal clearance between 0 and 20% were colored blue, and 45 compounds that have a percent renal clearance between 20 and 100% were colored red. The score plot in Figure 1 shows that compounds with a higher percent renal clearance (red) are present on the right side of the graph, whereas compounds with a low percent renal clearance (blue) are on the left side. Even though the two categories of compounds were not clearly separated, this score plot is significant because no biological data was given as input. The score plot just obtained from Volsurf descriptors calculated from the 3D structures of the highly diverse data set gave an idea about the renal clearance behavior of these compounds. The PCA model was validated by an external test set of 20 compounds (Table 1). In Volsurf, it is possible to apply a PCA model to an external test set, to obtain "predicted" scores that can be used to obtain score plots representing both the original series and the external objects. This representation can be seen as a projection of the external series in the same dimensionally reduced space obtained for the original series, with the rotation defined by the original loading matrix. Figure 2 shows the score plot of PC1 versus PC2 for the predicted scores of the test set. Here also, most of the red-colored compounds, which have a high percent of renal clearance, were present on the right side of the plot and the blue-colored compounds, which have a low clearance, were toward the left. Because a strong signal was obtained from PCA that there exists a correlation between Volsurf descriptors and renal clearance data, we conducted a PLS analysis where the percent renal clearance data were given as input. The PLS analysis yielded a four-component model with a cross-validated r^2 (q^2) value of 0.768 and a conventional r^2 value of 0.844 (Table 3). Figure 3 shows the score plot of PC1 versus PC2 for the PLS model, which appears clearer compared to the PCA model (Figure 1) as a result of training given in the form of percent renal clearance. This PLS model can be used as a projection map to understand the renal clearance of unknown compounds. The PLS model was also validated in the same way as that of the PCA model by using the same test set of 20 compounds. Figure 4 shows the score plot of PC1 versus PC2 for the predicted scores of the test set obtained from the PLS model. The model successfully predicted red-colored compounds on the right-hand side the plot and blue-colored compounds on the left.

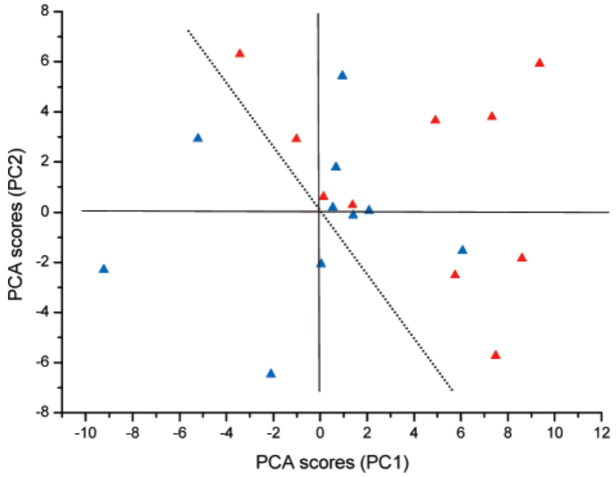

Figure 2. Score plot of PC1 vs PC2 for the PCA analysis. Test set compounds are colored according to percent renal clearance. Blue (0−20%) and red (20−100%).

Table 3. Summary of PLS Analysis (Volsurf)[a]

	% renal clearance
q^2	0.768
r^2	0.844
N	4
SDEP	13.43
SDEC	11.02

[a] q^2, cross-validated correlation coefficient; *N*, optimum number of components; r^2, non-cross-validated correlation coefficient; SDEP, standard deviation of error of predictions; SDEC, standard deviation of error of calculations.

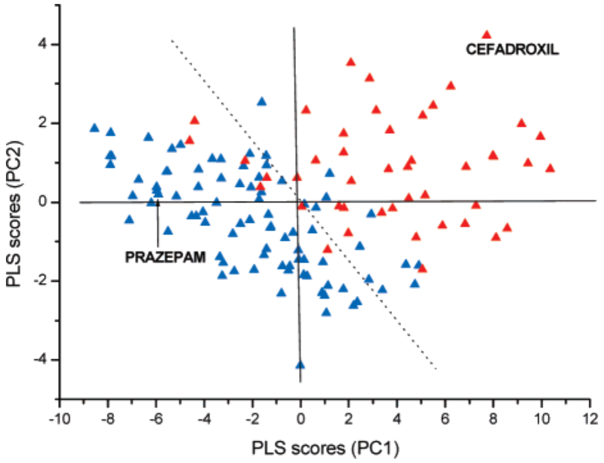

Figure 3. Score plot of PC1 vs PC2 for the PLS analysis. Training set compounds are colored according to percent renal clearance. Blue (0−20%) and red (20−100%). Also indicated are the positions of cefadroxil (93%) and prazepam (0%).

To understand the correlation of Volsurf descriptors with renal clearance, a PLS coefficient plot was drawn (Figure 5). The PLS coefficient plot shows that renal clearance is directly proportional to high values of hydrophilic regions (WOH2), best volumes (BVOH2), capacity factors (CwOH2), and local interaction energy minima (EminOH2) of the water probe, whereas it is inversely proportional to integy moments (IwOH2). In the case of descriptors generated from the

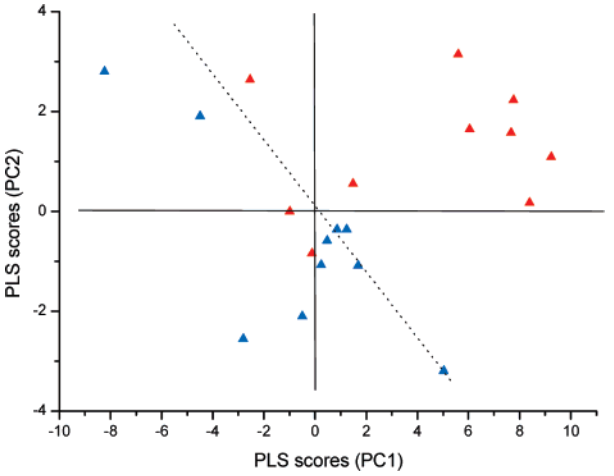

Figure 4. Score plot of PC1 vs PC2 for the PLS analysis. Test set compounds are colored according to percent renal clearance. Blue (0−20%) and red (20−100%).

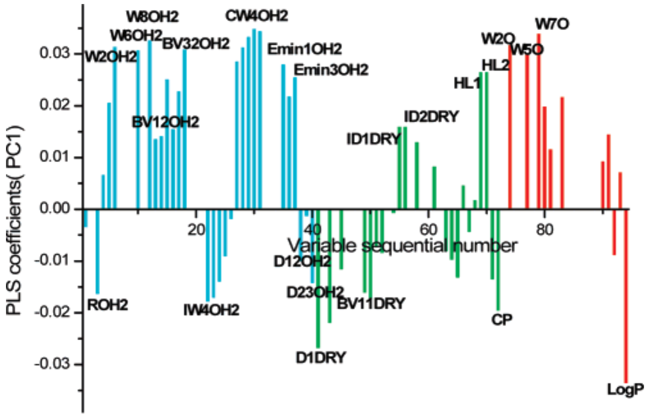

Figure 5. PLS coefficients plot for the correlation of Volsurf descriptors with renal clearance.

hydrophobic probe, percent renal clearance increases with higher values of hydrophobic integy moments (IDDRY) and hydrophilic−lipophilic balance (HL) and decreases with higher values of hydrophobic regions (DDRY), hydrophobic best volumes (BVDRY), amphiphilic moment (A), and critical packing (CP). Volumes of interactions (WO) and H-bond interaction energies (HBO) of the carbonyl probe are directly proportional to the percent renal clearance. But one of the most important descriptors, as observed from the PLS coefficient plot, is log *P* (water/octanol partition coefficient), which is inversely proportional to the percent renal clearance. In general, it is known that passive renal elimination is restricted to hydrophobic compounds. As a result, it may be expected that there is a relationship between the renal clearance and lipophilicity, the general trend of which is, as the log *P* value increases, the renal clearance decreases. Compounds which have a low renal clearance like caffeine (1.1%), warfarin (2%), and verapamil (3%) are shown to have a high percentage of oral absorption of about 100%, 93%, and 95%,[29,30] respectively. This suggests that processes such as renal clearance, which are governed by lipophilicity, are the key determinants for oral absorption and, in turn, bioavailability.

Figure 6 shows the Volsurf descriptor profile for one of the high renal clearance compounds, cefadroxil (93%), which is present on the extreme right side of the PLS score plot (Figure 3). The values of the descriptors are clearly in agreement with the proportionality shown in the PLS coefficient plot (Figure 5). But in the case of the Volsurf descriptor profile of prazepam (Figure 7), one of the compounds with a 0% renal clearance, which is present on the left side of the PLS score plot (Figure 3), the descriptor values are exactly opposite of that of the cefadroxil and PLS coefficient plots. These plots give an idea about the descriptors which are responsible for the peculiar behavior of these extreme compounds.

Figures 8 and 9 give a visual comparison of the Grid 3D molecular fields of cefadroxil and prazepam calculated with the water probe, respectively. The cyan regions around molecules show hydrophilic regions. The arrows represent the vectors of integy moments, which measure the unbalance between the center of mass and the position of the hydrophilic regions around them. Cefadroxil has larger hydrophilic regions compared to prazepam, whereas prazepam has larger integy moments compared to cefadroxil. Higher integy moments indicate that the hydrated regions are clearly concentrated in only one part of the molecular surface. Smaller integy moments indicate that the polar moieties are either close to the center of mass or at opposite ends of the molecule; thereby, the resulting barycenter is close to the center of the molecule. These 3D grid maps are in agreement with the PLS coefficient plot (Figure 5), where it is shown that percent renal clearance is directly proportional to hydrophilic regions (WOH2) and inversely proportional to integy moments (IW0H2). The PLS model was used to predict the percent renal clearance values of both the training set and the test set (Figure 10). Blue triangles indicate the training set, and red triangles indicate the test set. Table 1 shows the experimental versus calculated renal clearance data

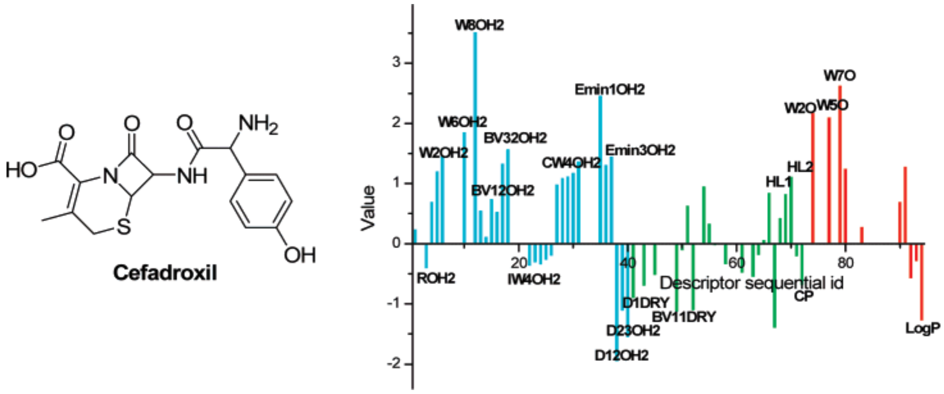

Figure 6. Volsurf descriptor profile of cefadroxil.

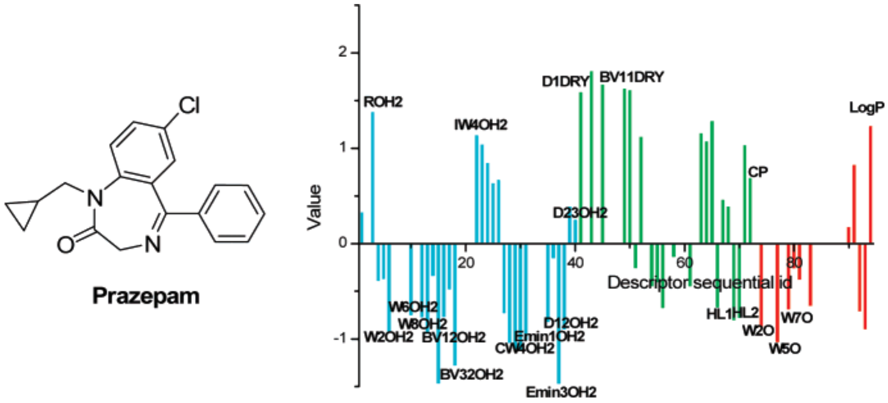

Figure 7. Volsurf descriptor profile of prazepam.

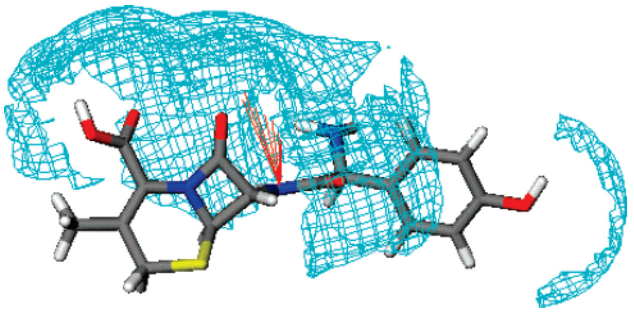

Figure 8. Grid 3D molecular fields of cefadroxil calculated with a water probe. The arrows represent the integy moment's pattern.

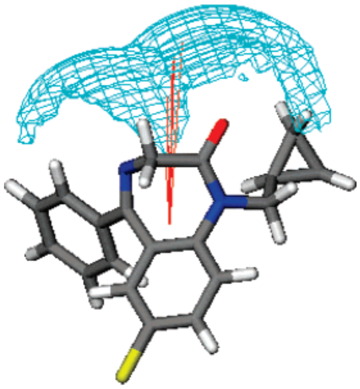

Figure 9. Grid 3D molecular fields of prazepam calculated with a water probe. The arrows represent the integy moment's pattern.

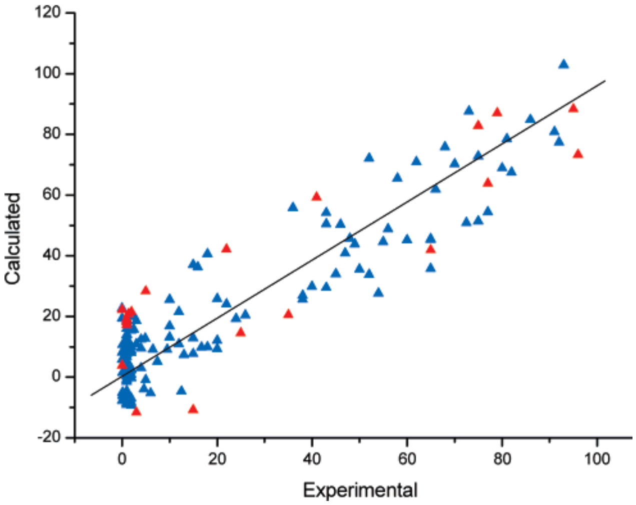

Figure 10. Plot of experimental vs calculated renal clearance for both the training (blue triangles) and test (red triangles) sets obtained from PLS analysis (Volsurf).

Table 4. Summary of PLS Analysis (Molconn-Z)[a]

	% renal clearance
q^2	0.530
r^2	0.720
N	4
SDEP	19.47
SDEC	15.05

[a] q^2, cross-validated correlation coefficient; N, optimum number of components; r^2, non-cross-validated correlation coefficient; SDEP, standard deviation of error of predictions; SDEC, standard deviation of error of calculations.

for both the training set and the test set. Most of the compounds were fairly predicted with residual values less than 20%.

For comparative purposes, we generated 95 Molconn-Z descriptors for the training set and studied the ability of this class of descriptors to predict renal clearance. Using a correlation matrix, 95 descriptors were filtered to 37 using a correlation threshold of 0.90. A PLS analysis of the 130-compound training set yielded a four-component model with a cross-validated r^2 (q^2) value of 0.530 and a conventional r^2 value of 0.720 (Table 4). This model was also used to predict the percent renal clearance of all the compounds of the data set (Table 1). Figure 11 shows the experimental versus calculated values of both the training and test sets. The q^2 and r^2 values and the predicted percent renal clearance data show that the Volsurf analysis showed better predictions than the Molconn-Z descriptors. For the purpose of classifying these compounds into low- and high-clearance compounds, categorical statistical techniques such as SIMCA and recursive partitioning were also examined. A total of 85 compounds that have a low renal clearance of less than 20% were grouped into class 1, and 45 compounds with a high percent of renal clearance of more than 20% were grouped into class 2. Both a SIMCA and a RP analysis were conducted to see whether these models correctly predict the compounds in their respective groups. Table 5 shows the summary of the SIMCA and RP analyses. Both Volsurf and Molconn-Z descriptors classified 80% of the training set compounds correctly in the case of SIMCA. But the predictive ability in the case of the test set was better in the case of the Molconn-Z descriptors, with 85% of the 20-compound test set predicted correctly compared to 65% by the Volsurf descriptors. Recursive partitioning classification was comparatively better than SIMCA, where 88% of the

Table 5. Summary of SIMCA and RP Analysis[a]

	Volsurf						Molconn-Z					
	train	cp	%cp	test	cp	%cp	train	cp	%cp	test	cp	%cp
						SIMCA						
class 1	85	68	80.00	10	5	50	85	72	84.71	10	9	90
class 2	45	34	75.55	10	8	80	45	35	77.78	10	8	80
total	130	102	78.46	20	13	65	130	107	82.31	20	17	85
						RP						
class 1	85	72	84.71	10	6	60	85	75	88.24	10	7	70
class 2	45	43	95.56	10	8	80	45	39	86.67	10	8	80
total	130	115	88.46	20	14	70	130	114	87.69	20	15	75

[a] train, training set; cp, correctly predicted; %cp, percentage correctly predicted; test, test set; class 1 (0−20%); class 2 (20−100%).

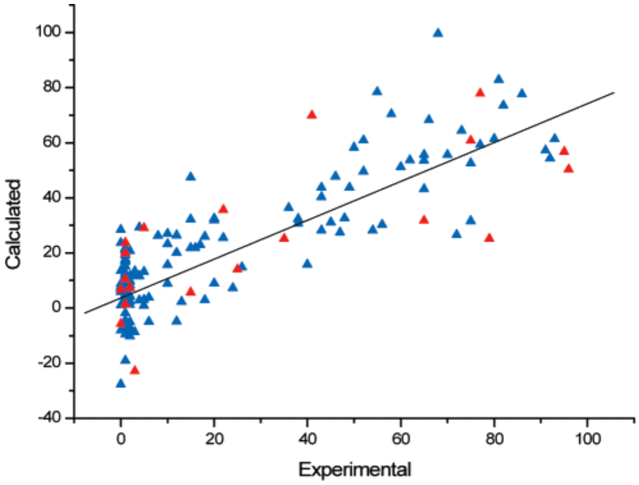

Figure 11. Plot of experimental vs calculated renal clearance for both the training (blue triangles) and test (red triangles) sets obtained from PLS analysis (Molconn-Z).

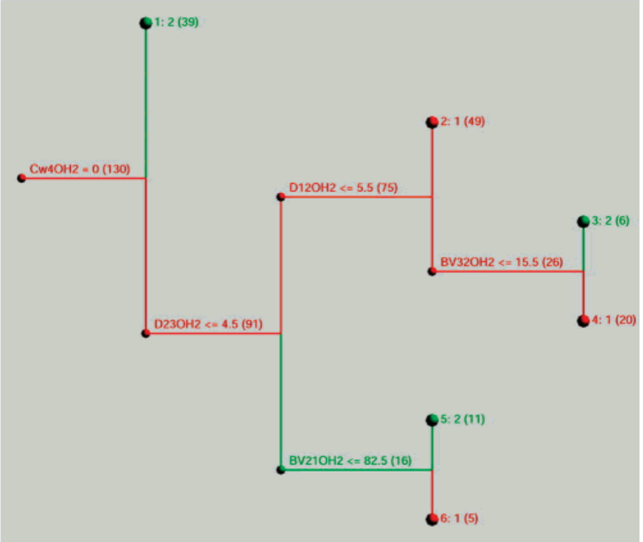

Figure 12. Decision tree obtained from RP of Volsurf descriptors. Red, class 1; green, class 2.

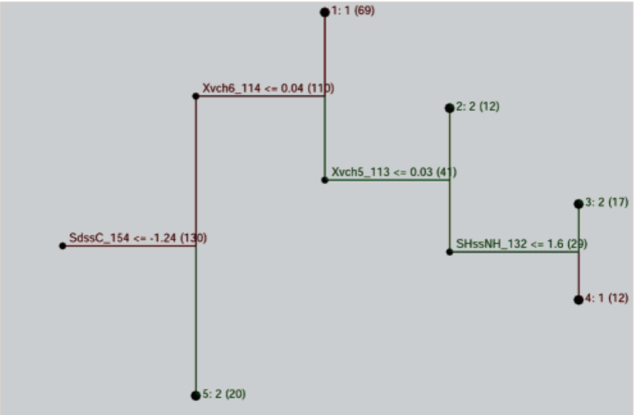

Figure 13. Decision tree obtained from RP of Molconn-Z descriptors. Brown, class 1; green, class 2.

training set compounds were correctly predicted in both cases. RP models were also validated by using the same 20-compound test set. More than 70% of the test set was correctly predicted in their respective classes by both the models. Figure 12 shows the decision tree obtained from the RP of the Volsurf descriptors. This decision tree is in agreement with the PLS coefficient plot (Figure 5). Cw4OH2, BV21OH2, and BV32OH2, whose higher values separated high-renal-clearance compounds (class 2) from low-clearance compounds, were shown to be directly proportional to renal clearance in the PLS plot. D23OH2 and D12OH2, whose higher values separated low-clearance compounds (class1) from high-clearance compounds, were shown to be inversely proportional to renal clearance in the PLS plot. Similarly, Figure 13 shows the decision tree obtained from RP of the Molconn-Z descriptors. SdssC, Xvch6, Xvch5, and SHssNH[15] were shown to be the important descriptors separating the two categories. Low-clearance compounds have comparatively high values for SdssC and Xvch6 and low values for SHssNH.

CONCLUSIONS

A renal clearance model was generated by using a classical Volsurf approach. The score plots obtained from PCA and PLS analyses clearly separated compounds with a higher renal clearance from compounds with a lower renal clearance. PLS models were used to predict the renal clearance of both training and test sets. SIMCA and RP techniques were used to classify the data set in to low- and high-clearance categories. For comparative purposes, topological descriptors such as Molconn-Z were also examined. The PLS model of the Volsurf descriptors showed better predictive ability compared to that of Molconn-Z. The PLS coefficient plot was used to explain the correlation of Volsurf descriptors with the renal clearance data. 3D Grid maps and a Volsurf descriptor profile of cefadroxil (high clearance) and prazepam (low clearance) were used to explain the structural and chemical features separating low- and high-clearance compounds. All of the models were validated by an external test set of 20 compounds. The successful classification and prediction of the renal clearance of a diverse set of compounds shows the efficiency of this kind of approach in understanding the bioavailability of druglike molecules.

ACKNOWLEDGMENT

This work was supported by Korea Institute of Science and Technology.

REFERENCES AND NOTES

(1) Norris, D. A.; Leesman, G. D.; Sinko, P. J.; Grass, G. M. Development of predictive pharmacokinetic simulation models for drug discovery. *J. Controlled Release* **2000**, *65*, 55.

(2) Grass, G. M.; Sinko, P. J. Effect of diverse datasets on the predictive capability of ADME models in drug discovery. *Drug Discovery Today* **2001**, *6*, S54.

(3) Ekins, S.; Waller, C. L.; Swaan, P. W.; Cruciani, G.; Wrighton, S. A.; Wikel, J. H. Progress in predicting human ADME parameters in silico. *J. Pharmacol. Toxicol. Meth*. **2000**, *44*, 251.

(4) Ghafourian, T.; Fooladi, S. The effect of structural QSAR parameters on skin penetration. *Int. J. Pharm*. **2001**, *217*, 1.

(5) Feher, M.; Sourial, E.; Schmidt, J. M. A simple model for the prediction of blood−brain partitioning. *Int. J. Pharm*. **2000**, *201*, 239.

(6) Martens, H.; Naes, T. *Multivariate Calibration*; Wiley: Chichester, U.K., 1998.

(7) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37.

(8) Dunn, W. J.; Wold, S. Pattern Recognition Techniques in Drug Design. In *Comprehensive Medicinal Chemistry*; Hansch, C., Sammes, P. G., Taylor, J. B., Eds.; Pergamon Press: Oxford, U. K., 1990; Vol. 4, pp 691−71.

(9) Turner, J. V.; Maddalena, D. J.; Cutler, D. J.; Agantonovic-Kustrin, S. Multiple pharmacokinetic parameter prediction for a series of cephalosporins. *J. Pharm. Sci*. **2003**, *92*, 518.

(10) Yoshida, F.; Topliss, J. G. QSAR model for drug human oral bioavailability. *J. Med. Chem*. **2000**, *43*, 2575.

(11) Andrews, C. W.; Bennett, L.; Yu, L. Y. Predicting human bioavailability of a compound: development of a novel quantitative structure−bioavailability relationship. *Pharm. Res*. **2000**, *17*, 639.

(12) Gaviraghi, G.; Barnaby, R. J.; Pellegatti, M. Phamacokinetic challenges in lead optimization. In *Pharmacokinetic Optimization in drug Research*; Testa, B., Van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Wiley-VCH: Basel, Switzerland, 2001; pp 1−14.

(13) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11* (2), S29.

(14) Cruciani, G.; Crivori, P.; Carrupt, P. A.; Testa, B. Molecular fields in quantitative structure−permeation relationships: the VolSurf approach. *THEOCHEM* **2000**, *503*, 17.

(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd/Wiley: Letchworth, Hertfordshire, England, 1986.

(16) Wold, S. Pattern recognition by means of disjoint principal component models. In *Pattern recognition*; 1976; Vol. 8, p 127.

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

(18) Crivori, P.; Cruciani, G.; Carrupt, P.; Testa, B. Predicting blood−brain permeation from three-dimensional molecular structure. *J. Med. Chem*. **2000**, *43*, 2204.

(19) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(20) Hardman, J. G.; Goddman, A. G.; Limbard, L. E. *Goddman and Gilman's The Pharmacological Basis of Therapeutics*; McGraw-Hill: New York, 1996; pp 1710−1792.

(21) *XConcord*, version 5.1.2; Optive Research, Inc.: Austin, TX, 1986−2003.

(22) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. The Tripos Force Field. *J. Comput. Chem.* **1989**, *10*, 982.

(23) *SYBYL*, version 7.0; Tripos Inc.: St. Louis, MO, 63144.

(24) *Volsurf*, version 4.1; Molecular Discovery Ltd.: Middlesex, U. K., 2000−2004.

(25) Cruciani, G.; Clementi, S. GOLPE: Philosophy and Applications in 3D-QSAR. In *Advanced Computer-Assisted Techniques in Drug Discovery*; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany, 1994; pp 61−88.

(26) Cruciani, G.; Clementi, S.; Baroni, M.; Pastor, M. Recent Development in 3D-QSAR Methodologies. In *Rational Molecular Design in Drug Research*; Alfred Benzon Symposium 42; Liljefors, T., Jorgensen, F. S., Krogsgaard-Larsen, P., Eds.; Munskgaard: Copenhagen, Denmark, 1998; pp 87−97.

(27) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Helberg, S.; Johansson, E.; Lindberg, W.; Sjostrom, M. Multivariate Data Analysis in Chemistry. In *Chemometrics Mathematics and Statistics in Chemistry*; Kowalsky, B. R., Ed.; Dordrecht: Holland, 1983; pp 17−96.

(28) *Cerius2*, version 4.10; Accelrys Inc.: San Diego, CA, 2005.

(29) Gres, M.; Julian, B.; Bourrie, M.; Meunier, V.; Riques, C.; Berger, M.; Boulenc, X.; Berger, Y.; Fabre, G. Correlation between oral drug absorption in humans, and apparent drug permeability in TC-7 Cells, A human epithelial intestinal cell line: comparison with the parental Caco-2 cell line. *Pharm. Res*. **1998**, *15*, 726.

(30) Kansy, M.; Senner, F.; Gubernator, K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. *J. Med. Chem*. **1998**, *41*, 1007.

CI0503309