Predicting Caco-2 Permeability Using Support Vector Machine and Chemistry Development Kit

Ma Guangli, Cheng Yiyu*

Pharmaceutical Informatics Institute, Zhejiang University, Hangzhou 310027, China.

Received, March 28, 2006; Accepted, July 11, 2006; Published, July 11, 2006.

ABSTRACT PURPOSE: To predict Caco-2 permeability is a valuable target for pharmaceutical research. Most of the Caco-2 prediction models are based on commercial or special software which limited their practical value. This study represents the relationship between Caco-2 permeability and molecular descriptors totally based on open source software. **METHODS:** The Caco-2 prediction model was constructed based on descriptors generated by open source software Chemistry Development Kit (CDK) and a support vector machine (SVM) method. Number of H-bond donors and three molecular surface area descriptors constructed the prediction model. **RESULTS:** The correlation coefficients (*r*) of the experimental and predicted Caco-2 apparent permeability for the training set and the test set were 0.88 and 0.85, respectively. **CONCLUSIONS:** The results suggest that the SVM method is effective for predicting Caco-2 permeability. Membrane permeability of compounds is determined by number of H-bond donors and molecular surface area properties.

INTRODUCTION

Over the past 10 years, much attention has been paid to absorption, distribution, metabolism, and elimination (ADME) screening, because of the important role of ADME screening in modern drug development. Many *in vitro* ADME screening methods have been applied to boost drug discovery process in pharmaceutical industry. Although *in vitro* ADME screening methods are high performance compared with *in vivo* ADME screening protocols, they are still resource-intensive and time-consuming.

Corresponding author: Dr. Cheng Yiyu, Telephone: 86-571-879-51138; Fax: 86-571-879-51138; E-mail: chengyy@zju.edu.cn.

To further improve ADME screening, various *in silico* screening methods (ADME *in silico*) have been constructed (1-3), for e.g. human oral absorption (4-7), bioavailability (8, 9), metabolism (10, 11), P-glycoprotein substrates (12, 13).

Since oral is the most favorite way in various routines for drug delivery, estimating human oral bioavailability of candidates in the early stage of the drug development process is important and necessary for lead selection and optimization. Screening for absorption ability is an important part of assessing oral bioavailability and attracts efforts from industry and academia. In several *in vitro* cell culture models for drug absorption, the cell line most widely used is Caco-2 cells (14). These are well-differentiated intestinal cells derived from human colorectal carcinoma. These cells retain many morphological and functional properties of the *in vivo* intestinal epithelial cell barrier, which makes the Caco-2 cell monolayer an important model for *in vitro* absorption screening. Extensive studies have revealed that the human oral absorption of compounds is related to Caco-2 permeability (15). Thus, Caco-2 permeability is a valuable index for assessing oral absorption of compounds, which, in turn, calls for the methods for predicting chemical Caco-2 permeability.

Research on predicting Caco-2 permeability from structures of compounds using quantitative structure property relationship (QSPR) modeling is on the way. Some studies are summarized in Table 1. In these studies, various types of molecular descriptors were employed such as dynamic polar surface area (PSAd), HBA, HBD, MW, logP, logD, high charged polar surface area (HCPSA), radius of gyration (rgyr), RB, and membrane-interaction descriptors. Most of the Caco-2 permeability prediction models were based on linear methods such as linear regression, multiple linear regression (MLR), or partial least squares (PLS). They generally used small sets of molecules and were not fully validated by external test sets. Fujiwara *et al* introduced neural networks to enhance regression ability of Caco-2 permeability prediction models (21). Given the common recognition that statistical significance of a QSAR/QSPR model does not imply its practical applicability, a validation using external test sets of molecules is necessary. Hou *et al* collected several published data sets and

investigated the relationship between simple descriptors and Caco-2 permeability(23). Most of the models were built by commercial or special software packages such as SYBYL(23, 25) and VolSurf (11), which limited the usage and validation of the models by other researchers.

In this study, Caco-2 permeability prediction models based on MLR and SVM methods were built. All of the molecular descriptors involved were calculated by open source software, CDK and statistical work was done by open source software, R. Open source means that the software is free and the users can read, modify and add functions to the source code freely. This means those who want to validate or use the models described here to predict Caco-2 permeability from the chemical structure can do so. Furthermore, the models provided some insight into physiochemical process. Because Caco-2 penetration mentioned here was passive transport, metabolism and active transport was out of the scope of this study.

MATERIALS AND METHODS

Software and Hardware

The chemistry development kit (CDK) is a freely available open source Java library for structural chemistry and bioinformatics. Its development is an open source project by a team of international collaborators from academic and industrial institutions. CDK provides methods for many tasks in molecular informatics, such as 2D and 3D rendering of chemical structure, I/O routines, SMILES parsing and generation, ring searches, isomorphism checking, structure diagram generation, and energy minimization (26). In the recent update, CDK provides QSAR modeling functions which included more than 30 routines to calculate descriptors and an interface to open source statistical software, R (27).

If the software involved in QSAR/QSPR modeling is commercial, the validation and usage of the models is limited for the other researchers. The CDK project provides not only source code but also references to dictionaries that describe the exact algorithm used, and versioning of the implementation of the algorithm, which makes academic research more valuable to both academics and industry (28).

The descriptors studied in this work were generated by CDK version 20050826. The CDK and R were linked by SJava 0.8 on Linux Redhat Workstation 4.0, Compaq Evo N600C. The statistical methods were provided by R version 2.2.0. Data manipulation and model scripts in R were written.

Data Set

The experimental apparent Caco-2 permeability($logP_{app}$) data of 100 drugs was collected from the literature (23). To compare with Hou *et al*'s work, the data set was separated into a training set of 77 compounds and a test set of 23 compounds as in Hou *et al* (23). The training set was used to build model, and the test set was used to evaluate its predictability. The experimental apparent permeability, the molecular descriptors used in this study and the results predicted by the multiple regression model (MLR) and the support vector machine (SVM) model were listed in Table 2.

Descriptors

Descriptors play a vital role in QSAR/QSPR models. Simple and meaningful descriptors make QSAR/QSPR models understandable and useful for drug development. The descriptors available in the current version CDK are separated into 5 classes: constitutional, topological, geometric, electric, and hybrid (28) as shown in Table 3. Further information about the descriptors and CDK can be found on the web site (http://cdk.sf.net). That CDK provides a wide range of descriptors makes predicting Caco-2 permeability possible.

Correlation between membrane permeability and some descriptors was impossible or difficult to be understood and explained. These descriptors were removed from the data set. They were Chi series, eccentric connectivity, Kier value, Petitjean number, VAdjMa, Winer number, Zagre index, geometrical descriptors, BCUT class and WHIM class descriptors.

COMPUTATIONAL METHODS

Genetic algorithm (GA)

As a stochastic search technique, genetic algorithm (GA) based on the principle of natural

evolution, was widely used in pharmaceutical, chemical and bioinformatical investigations (3, 29-33). Haupt *et al* explained this technique from a practical point of view (34). Genetic algorithms are categorized into binary and continuous classes.

Binary GA used to select descriptors was extensively studied in recent investigations (35, 36). Continuous GA was employed to determine the optimal SVM parameters (37). In this paper, Binary GA was used to select variables from descriptors generated by CDK to build MLR and SVM models, respectively. Continuous GA optimized the SVM parameters to achieve best Caco-2 permeability prediction performance.

The core of the optimization problem is the evaluation function. The evaluation functions for descriptors selection were given below arbitrarily.

$$\text{MLR: } eva = 2 - r_{training} - 0.1 r_{test} + 0.01N$$

$$\text{SVM: } eva = 2 - r_{training} - r_{test} + 0.015N$$

where *eva* is the evaluation value. $r_{training}$ is the correlation coefficient of experimental and predicted values for training set. r_{test} is the correlation coefficient of experimental and predicted values for test set. N is number of descriptors. 0.01 N and 0.015 N are penalty components to reduce the number of descriptors selected into the models.

Table 1: Summary of Several Caco-2 Permeability Prediction Investigations

Year	Authors	Method	Software	Descriptors
1996	Palm *et al*(16)	Linear Regression	PCMODEL, MacroModel	Dynamic polar surface area (PSAd)
1997	Norinder *et al*(17)	PLS	MolSurf	Surface, logP, Polarity, HBAo[a], HBD[b], HBA, HBD, etc.
1998	Camenisch *et al*(18)	Non-linear Regression	Statistica (statistical software)	MW, logD(oct)[c]
2000	Pickett *et al*(19)	Not Mentioned	Chem-X, SYBYL	ClogP[d], MW, PSA[e]
2000	Cruciani *et al*(11)	PLS	VolSurf	VolSurf Descriptors
2002	Kulkarni *et al*(20)	membrane-interaction QSAR(MI-QSAR)	Chemlab-II, Mopac 6.0	Solute aqueous dissolution and salvation descriptors, Solute-membrane interaction and salvation descriptors, General intramolecular solute descriptors. (Many descriptors)
2002	Fujiwara *et al*(21)	Molecular orbital (MO) calculation, 5-4-1 BP neural network	MOPAC97	Dipole moment, Polarizability, Sum(N)[f], Sum(O)[g], Sum(H)[h]
2002	Yamashita *et al*(3)	Genetic Algorithm Based Partial Least Squares	Molconn-Z 3.50	Molconn-Z descriptors
2003	Ponce *et al*(22)	MLR	TOMO-COMD	Quadratic Indices
2004	Hou *et al*(23)	MLR	SYBYL, SASA, MSMS, etc.	HCPSA[i], logD, rgyr[j], RB
2004	Ponce *et al*(24)	Linear discriminant analysis (LDA)	TOMO-COMDStatistica 5.5	Quadratic Indices
2005	Refsgaard *et al*(25)	Nearest-Neighbor classification	SYBYL, Matlab	Number of flex bonds, number of hydrogen bond acceptors and donors, molecular and polar surface area

[a] Hydrogen bond acceptor strength for oxygen atoms. [b] Hydrogen bond donor strength. [c] distribution coefficient in 1-octanol/water. [d] Calculated logP. [e] Polar surface area. [f] Sum of charges of nitrogen atoms. [g] Sum of charges of oxygen atoms. [h] Hydrogen atoms bonding to nitrogen or oxygen atoms. [i] High charged polar surface area. [j] Radius of gyration.

Table 2: Data of Experimental Apparent Permeability, Molecular Descriptors, and Predicted Results

Class	Name	Descriptors							Caco-2	Prediction	
		MLR				SVM					
		CPSA10	CPSA20	TPSA	HBD	CPSA0	CPSA18	CPSA27		MLR	SVM
Tr	acebutolol	1.50	-36.04	87.66	3.00	596.62	223.11	0.82	-5.83	-5.35	-5.75
Tr	Acebutolol_ester	1.81	-32.55	93.73	2.00	658.92	197.50	0.84	-4.61	-5.17	-4.69
Tr	acetylsalicylic_acid	1.15	-27.50	89.90	1.00	332.43	170.12	0.69	-5.06	-5.02	-4.98
Tr	acyclovir	1.47	-25.72	109.83	3.00	372.31	90.92	0.64	-6.15	-5.44	-5.54
Tr	alprenolol	1.11	-13.98	41.49	2.00	453.12	118.57	0.91	-4.62	-4.92	-4.95
Tr	Alprenolol_ester	1.24	-17.83	47.56	1.00	493.40	171.13	0.92	-4.47	-4.77	-4.55
Tr	aminopyrine	0.91	-6.54	26.79	0.00	404.16	66.44	0.93	-4.44	-4.49	-4.52
Tr	artemisinin	1.16	-13.22	53.99	0.00	355.86	61.90	0.81	-4.52	-4.62	-4.44
Tr	artesunate	1.75	-27.69	100.52	1.00	440.08	118.91	0.72	-5.40	-5.02	-5.15
Tr	atenolol	1.34	-19.12	84.58	3.00	415.72	102.17	0.80	-6.44	-5.31	-5.51
Tr	betazolol	1.30	-21.67	50.72	2.00	569.95	200.16	0.92	-4.81	-4.97	-4.88
Tr	betazolol_ester	1.53	-27.35	56.79	1.00	696.18	260.24	0.93	-4.52	-4.81	-4.60
Tr	bremazocine	1.12	-15.05	43.70	2.00	410.63	88.20	0.81	-5.10	-4.93	-5.02
Tr	caffeine	0.96	-14.53	53.51	0.00	338.12	65.17	0.71	-4.41	-4.64	-4.49
Tr	chloramphenicol	1.15	-32.02	112.70	3.00	279.76	134.88	0.68	-4.69	-5.50	-4.77
Tr	chlorothiazide	8.73	-748.62	135.45	2.00	90.61	224.82	0.50	-6.72	-6.57	-6.64
Tr	chlorpromazine	0.65	-10.92	31.78	0.00	369.64	156.40	0.98	-4.70	-4.55	-4.66
Tr	cimetidine	1.13	-19.53	98.40	3.00	462.23	112.32	0.92	-5.89	-5.40	-5.57
Tr	clonidine	0.63	-9.61	36.42	2.00	295.02	98.10	0.94	-4.59	-4.93	-4.66
Tr	corticosterone	1.38	-24.50	74.60	2.00	442.37	95.81	0.73	-4.47	-5.09	-5.09
Tr	desipramine	0.87	-5.74	15.27	1.00	433.37	81.51	0.98	-4.67	-4.61	-4.59
Tr	dexamethasone	1.63	-27.83	94.83	3.00	371.91	105.40	0.69	-4.75	-5.36	-5.16
Tr	dexamethasone_b_D_glucoside	2.79	-64.10	173.98	6.00	535.37	239.90	0.61	-6.54	-6.28	-6.48
Tr	dexamethasone_b_D_glucuronide	2.68	-73.07	191.05	6.00	497.74	279.22	0.55	-6.12	-6.40	-6.04
Tr	diazepam	0.71	-12.75	32.67	0.00	296.56	120.04	0.90	-4.32	-4.55	-4.40
Tr	dopamine	0.73	-16.36	66.48	3.00	288.12	93.00	0.72	-5.03	-5.27	-5.11
Tr	doxorubicin	2.92	-57.34	206.07	6.00	541.60	200.98	0.64	-6.80	-6.41	-6.72
Tr	erythromycin	4.24	-46.38	193.91	5.00	745.77	189.21	0.82	-5.43	-6.02	-5.51
Tr	estradiol	0.98	-12.58	40.46	2.00	385.05	68.88	0.82	-4.77	-4.92	-4.77
Tr	felodipine	1.25	-16.27	64.63	1.00	430.69	130.42	0.90	-4.64	-4.85	-4.60
Tr	ganciclovir	1.69	-32.33	130.06	4.00	386.35	109.90	0.62	-6.27	-5.72	-6.20
Tr	griseofulvin	1.42	-25.03	71.06	0.00	395.24	107.21	0.79	-4.44	-4.71	-4.52
Tr	Hydrochlorothiazide	12.70	-788.01	135.12	3.00	137.45	222.37	0.50	-6.06	-6.47	-6.14

Table 2 continued…

Tr	hydrocortisone	1.56	-25.71	94.83	3.00	399.23	91.94	0.69	-4.66	-5.36	-5.32
Tr	ibuprophen	0.78	-12.30	37.30	1.00	363.39	92.04	0.79	-4.28	-4.75	-4.63
Tr	imipramine	0.88	-4.09	6.48	0.00	434.31	76.43	1.00	-4.85	-4.38	-4.77
Tr	indomethacin	1.00	-27.35	63.60	1.00	357.09	183.88	0.77	-4.69	-4.90	-4.77
Tr	labetalol	1.45	-25.03	95.58	4.00	427.99	144.18	0.80	-5.03	-5.55	-5.29
Tr	mannitol	1.38	-27.03	121.38	6.00	244.85	69.52	0.32	-6.21	-6.05	-6.13
Tr	meloxicam	15.00	-273.11	107.98	2.00	325.47	140.83	0.69	-4.71	-4.61	-4.64
Tr	methanol	0.33	-14.24	20.23	1.00	360.94	35.75	0.67	-4.58	-4.71	-4.65
Tr	methotrexate	2.24	-59.15	210.54	5.00	509.54	257.56	0.61	-5.92	-6.32	-6.00
Tr	Methylsco polamine	2.20	-18.51	59.06	1.00	421.30	84.86	0.84	-6.16	-4.74	-4.68
Tr	metoprolol	1.39	-15.83	50.72	2.00	528.62	92.25	0.89	-4.59	-4.95	-4.82
Tr	nadolol	1.67	-23.51	81.95	4.00	517.83	102.30	0.78	-5.41	-5.46	-5.49
Tr	naproxen	0.78	-17.80	46.53	1.00	319.28	122.13	0.78	-4.83	-4.81	-4.74
Tr	nevirapine	1.03	-10.74	58.12	1.00	386.36	84.98	0.86	-4.52	-4.83	-4.56
Tr	nicotine	0.55	-5.04	16.13	0.00	317.41	49.26	0.92	-4.71	-4.46	-4.78
Tr	olsalazine	1.14	-48.13	139.78	4.00	323.20	230.58	0.47	-6.96	-5.86	-6.88
Tr	oxprenolol	1.33	-13.56	50.72	2.00	461.71	109.12	0.91	-4.68	-4.95	-4.93
Tr	oxprenolol_ester	1.42	-17.98	56.79	1.00	476.90	156.06	0.91	-4.52	-4.80	-4.57
Tr	phencyclidine	0.80	-2.41	3.24	0.00	428.76	47.59	1.00	-4.61	-4.37	-4.68
Tr	phenytoin	0.86	-20.34	58.20	2.00	304.88	139.80	0.77	-4.57	-5.05	-4.65
Tr	pindolol	1.18	-12.44	41.49	3.00	430.73	88.44	0.87	-4.78	-5.09	-5.41
Tr	pirenzepine	1.52	-17.63	68.78	1.00	467.10	109.05	0.85	-6.36	-4.85	-4.85
Tr	piroxicam	15.50	-273.80	107.98	2.00	341.27	139.42	0.66	-4.45	-4.56	-4.64
Tr	pnu200603	1.07	-17.87	75.65	3.00	421.44	123.38	0.85	-6.25	-5.29	-5.67
Tr	practolol	1.32	-18.31	70.59	3.00	472.52	110.33	0.83	-6.05	-5.24	-5.81
Tr	prazocin	1.90	-25.83	93.81	1.00	521.95	123.02	0.80	-4.36	-4.96	-4.99
Tr	progesterone	0.93	-13.68	34.14	0.00	436.42	74.31	0.87	-4.37	-4.54	-4.44
Tr	propranolol	1.12	-10.23	41.49	2.00	426.58	90.59	0.91	-4.58	-4.91	-4.75
Tr	propranolol _este	1.24	-14.31	47.56	1.00	483.75	153.09	0.94	-4.48	-4.76	-4.52
Tr	quinidine	1.31	-13.68	45.59	1.00	429.72	72.12	0.85	-4.69	-4.74	-4.62
Tr	ranitidine	1.52	-34.47	95.74	2.00	549.83	126.35	0.86	-6.31	-5.21	-5.24
Tr	Salicylic _acid	0.65	-24.43	57.53	2.00	305.47	127.09	0.53	-4.79	-5.07	-4.87
Tr	scopolamine	1.49	-17.82	62.30	1.00	443.56	89.23	0.80	-4.93	-4.82	-4.85
Tr	sucrose	2.51	-46.45	189.53	8.00	344.42	121.33	0.37	-5.77	-6.69	-5.69
Tr	sulphasalazine	12.28	-512.67	149.69	3.00	348.38	313.77	0.63	-6.33	-5.87	-6.25
Tr	telmisartan	1.69	-24.72	63.08	1.00	595.81	236.39	0.89	-4.82	-4.82	-4.74
Tr	terbutaline	1.15	-18.67	72.72	4.00	355.50	73.88	0.68	-6.38	-5.45	-6.30
Tr	tesosterone	0.98	-12.10	37.30	1.00	414.97	62.66	0.83	-4.34	-4.73	-4.57

Table 2 continued…

Tr	timolol	1.72	-18.40	79.74	2.00	489.33	70.47	0.87	-4.85	-5.07	-4.77
Tr	timolol_ester	1.99	-13.20	85.81	1.00	543.53	71.55	0.93	-4.60	-4.88	-4.54
Tr	uracil	0.53	-10.83	58.20	2.00	191.64	50.31	0.48	-5.37	-5.05	-5.41
Tr	urea	0.46	-10.21	69.11	2.00	182.08	39.24	0.50	-5.34	-5.11	-5.42
Tr	warfarin	1.05	-20.39	63.60	1.00	337.58	122.67	0.77	-4.68	-4.88	-4.75
Tr	zidovudine	1.26	-27.09	91.23	2.00	300.26	93.75	0.61	-5.16	-5.19	-5.08
Te	furosemide	11.29	-413.04	117.87	3.00	255.17	202.77	0.59	-6.50	-5.55	-5.66
Te	guanabenz	0.62	-15.50	74.26	3.00	250.54	122.51	0.87	-4.50	-5.32	-4.86
Te	fleroxacin	1.51	-32.50	64.09	1.00	404.23	144.12	0.61	-4.81	-4.87	-5.07
Te	mibefradil	1.92	-27.11	51.66	1.00	663.60	201.17	0.89	-4.87	-4.75	-4.38
Te	verapamil	1.86	-29.13	63.95	0.00	630.46	153.85	0.92	-4.58	-4.65	-4.81
Te	guanoxan	1.08	-16.56	80.36	3.00	348.13	101.62	0.84	-4.71	-5.31	-5.13
Te	saquinavir	3.05	-39.26	166.75	5.00	773.17	283.78	0.89	-6.26	-5.97	-5.47
Te	lidocaine	0.93	-6.35	32.34	1.00	400.38	65.89	0.95	-4.21	-4.69	-4.55
Te	enalapril	1.79	-28.05	95.94	2.00	523.33	158.14	0.77	-5.64	-5.17	-5.47
Te	theophylline	0.93	-18.51	53.51	1.00	336.07	76.96	0.64	-4.35	-4.83	-4.74
Te	cyclosporine	6.14	-73.97	278.80	5.00	1228.88	472.46	0.90	-6.05	-6.34	-5.46
Te	antipyrine	0.60	-7.06	23.55	0.00	275.73	59.22	0.87	-4.55	-4.50	-4.84
Te	proscillaridin	2.45	-60.89	125.68	4.00	626.80	226.92	0.68	-6.20	-5.70	-5.95
Te	coumarin	0.45	-11.58	26.30	0.00	264.25	89.41	0.80	-4.11	-4.54	-4.74
Te	nitrendipine	1.54	-31.77	107.77	1.00	432.61	119.21	0.78	-4.77	-5.08	-5.08
Te	epinephrine	0.97	-20.04	72.72	4.00	281.79	77.14	0.63	-6.02	-5.47	-6.29
Te	tiacrilast	0.76	-24.51	95.27	1.00	261.49	139.44	0.66	-4.90	-5.08	-5.13
Te	amoxicillin	1.62	-36.74	158.26	4.00	382.56	163.24	0.60	-6.10	-5.88	-5.78
Te	diltiazem	1.60	-21.87	84.38	0.00	490.68	139.31	0.87	-4.38	-4.76	-4.61
Te	remikiren	20.82	-419.04	154.07	5.00	718.31	252.35	0.84	-6.13	-5.20	-5.60
Te	sulpiride	14.80	-560.13	110.11	2.00	395.39	219.24	0.70	-6.16	-5.38	-5.16
Te	bosentan	18.83	-385.70	154.03	2.00	576.72	232.99	0.76	-5.98	-4.77	-5.46
Te	ceftriaxone	4.51	-76.90	253.62	2.00	424.95	320.36	0.54	-6.88	-5.84	-5.85

Tr is training set; Te is test set.

Table 3: Descriptors Implemented by CDK Version 20050826

CDK version 20050826				
Topological	Geometrical	Constitutional	Hybrid	Electronic
Chi0, Chi0C, Chi0v, Chi0vC, Chi1, Chi1C, Chi1v, Chi1vC, EccentricConnectivity, KierValues, PetitjeanNumber, TPSA, VAdjMa, WienerNumber, ZagrebIndex	GravitationalIndex, MomentOfInertia	Apol, AromaticAtomsCount, AromaticBondsCount, Bpol, Lipinskifailures, RotatableBoundsCount, XlogP	BCUT, CPSA, WHIM	HBondDonors, HBondAcceptors

Support vector machine (SVM)

In addition to descriptor selection, computational method selection is another critical step for QSAR/QSPR modeling. MLR, PLS, Nearest-Neighbor classification and neural networks were utilized to address Caco-2 permeability prediction as shown in Table 1. Support vector machine (SVM), developed by Vapnik, as a novel type of machine learning (38), was widely used to solve classification and regression problems on human oral absorption (39), solubility (40), P-glycoprotein substrate s(41), blood-brain barrier penetrating and nonpenetrating agents (13), and metabolism (42). SVM was employed to fit a non-linear relationship between the Caco-2 permeability and the CDK descriptors.

The regression performances of SVM depend on several factors: cost of constraints violation (cost), epsilon in the insensitive-loss function (epsilon), the kernel type and its parameters. A Radial Basis Function (RBF) was chosen to be the kernel function because it is widely used in regression problems. The three parameters, cost, epsilon, and gamma, used to construct the final Caco-2 permeability prediction model were 3.37, 0.1, and 0.61.

RESULTS AND DISCUSSION

The CDK descriptors selected by GA to construct MLR and SVM model were listed in Table 4. That HBD appeared in both the MLR model and the SVM model suggested that HBD is an important factor to membrane permeability. This conclusion is supported by the references . The relationship between membrane permeability and molecular surface area properties especially PSA_d, were extensively studied (16, 17, 19, 23, 25). Topological polar surface area (TPSA) as a form of PSA_d was selected into the MLR model. CDK provided a class of descriptors, CPSA, to describe charged partial surface area properties. In both the MLR model and the SVM model, CPSA descriptors played vital roles to predict Caco-2 permeability presented by Table 4.

Multiple linear regression (MLR) Model

A multiple linear regression (MLR) model was built to predict Caco-2 permeability with four CDK descriptors:

$$logP_{app} = -0.18HBD + 0.095CPSA10 + 0.0026CPSA20 - 0.0051TPSA - 4.42$$

Table 4: Symbols and Explanation of Descriptors to Construct MLR and SVM models

Model	Symbols	Meaning
MLR	CPSA10	Partial positive surface area*total positive charge on the molecule/total molecular surface area
	CPSA20	Charge weighted partial negative surface area*total molecular surface area/1000
	TPSA	Topological polar surface area based on fragment contributions
	HBD	Number of H-bond donors
SVM	CPSA0	Sum of surface area on positive parts of molecule
	CPSA18	Partial negative surface area* total molecular surface area/1000
	CPSA27	Sum of solvent accessible surface areas of atoms with absolute value of partial charges less than 0.2/total molecular surface area

Table 5: Distribution of Experimental Apparent Permeability and Each Molecular Descriptor

		Descriptors							
Class		Caco2	CPSA10	CPSA20	TPSA	HBD	CPSA0	CPSA18	CPSA27
Training	Min	-6.96	0.33	-788.01	3.24	0.00	90.61	35.75	0.32
	Mean	-5.14	2.09	-54.45	78.28	2.12	411.24	125.95	0.77
	Max	-4.28	15.50	-2.41	210.54	8.00	745.77	313.77	1.00
Test	Min	-6.88	0.45	-560.13	23.55	0.00	250.54	59.22	0.54
	Mean	-5.33	4.35	-102.44	106.32	2.17	475.88	174.87	0.77
	Max	-4.11	20.82	-6.35	278.80	5.00	1228.88	472.46	0.95

Table 5 summarizes the distribution of the descriptors calculated by CDK for the compounds used in this study. This dataset consists of compounds that were very different in structure and membrane permeability. Each descriptor of the compounds covered a wide range.

Table 6: Correlation Coefficients (*r*) between Experimental Apparent Permeability and Each MLR Descriptor for The Training Set

	HBD	CPSA10	CPSA20	TPSA
Caco2	-0.65	-0.19	0.33	-0.65
HBD		0.16	-0.13	0.81
CPSA10			-0.81	0.41
CPSA20				-0.37

The Hou's MLR model based on fraction of rotatable bonds (frotb), logD, high charged polar surface area (HCPSA) and radius of gyration (rgyr) was rebuilt, while the limitation of logD<2.0 was cut out (23). The *r* of Hou's model in the training set and the test set were 0.81 and 0.70. The *r* showed that the four descriptors generated by CDK had equivalent regression and prediction ability as Hou's four descriptors without the limitation of logD<2.0.

Support Vector Machine (SVM) Model

MLR did not give the satisfied results as shown in Figure 1. It was believed that there exist nonlinear relationship between Caco-2 permeability and CDK descriptors. Hence, SVM method as a good nonlinear regression algorithm, was employed. The correlation coefficients between Caco-2 permeability and each descriptor used in SVM model were listed in Table 7. The relationship between experimental $logP_{app}$ and predicted values might be nonlinear. The results of Spearman test was also support the SVM model as listed in Table 8.

Table 7: Correlation Coefficients (r) between Experimental Apparent Permeability and Each SVM Descriptor for The Training Set

	HBD	CPSA0	CPSA18	CPSA27
Caco2	-0.65	0.07	-0.37	0.52
HBD		0.08	0.37	-0.62
CPSA0			0.28	0.51
CPSA18				-0.20

Table 8: Summary of Spearman Test Results for Training and Test Data Set

Class		Valid N	Spearman R	p-level
Training	MLR	77.00	0.71	0.00
	SVM	77.00	0.85	0.00
Test	MLR	23.00	0.78	0.00
	SVM	23.00	0.85	0.00

The correlation coefficients (*r*) of experimental $logP_{app}$ and predicted values by the SVM model were 0.88 and 0.85 for the training set and the test set, respectively. The SVM model was the final model. Figure 1 shows that while experimental $logP_{app}$<-5, the values predicted by the MLR and SVM models were larger than the experimental values.

The SVM method was also applied to Hou's four descriptors. The correlation coefficients (*r*) of experimental $logP_{app}$ and predicted values for the training set and the test set were 0.92 and 0.77, respectively. The regression ability of Hou's four descriptors was increased remarkably, which proved SVM's regression ability. This also proved that the four descriptors generated by CDK were

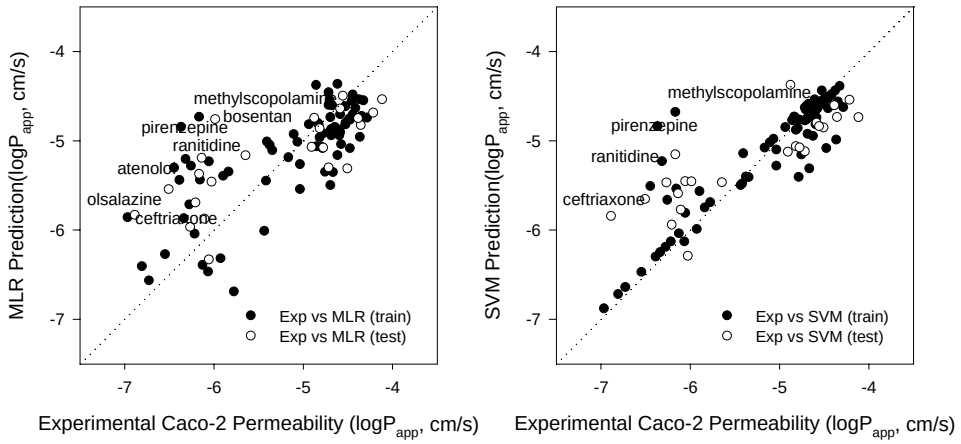

Figure 1: Correlation between experimental Caco-2 permeability and the predicted values by the MLR and SVM models.

equivalent compared with Hou's four descriptors for Caco-2 permeability prediction, when a non-linear modeling technique is used.

Most of the Caco-2 permeability prediction investigations were based on a small training set of molecules and no external test set but Hou *et al*(23) and Refsgaard *et al* (25). Refsgaard *et al* has built a classification model based on in-house data which could not be compared with this study. Because Hou's data set used in this study, it was convenient to compare with the results of Hou's work and Ponce's work mentioned in the literature (23). The results of the comparison are listed in Table 9. Unsigned mean error, *UME*, we calculated from Hou's data was 0.45, not 0.49. According to the comparison results, the SVM model based on CDK descriptors were significantly better than Hou's and Ponce's model significantly.

CONCLUSION

In the current study, Caco-2 permeability prediction models based on MLR, SVM and CDK descriptors were developed. The previous investigations on Caco-2 permeability prediction and the MLR model described in this paper suggested the nonlinear relationship between descriptors and Caco-2 permeability. The SVM method assigned CDK descriptors nonlinear regression ability to achieve good performance in Caco-2 permeability prediction, which implies that SVM method is an effective algorithm for Caco-2 permeability prediction. The descriptors selected in the MLR or SVM model represent that Caco-2 or membrane permeability is determined by number of H-bond donors and molecular surface area properties.

At last, this SVM model is not perfect, because the data set and descriptors used here were limited. A larger data set could make the model better for prediction and cover larger chemical space. That the whole work was based on open source software CDK makes everybody in pharmaceutical area free to use, rebuild models and even develop QSAR/QSPR software. The modeling investigation of oral absorption, distribution, and clearance prediction based on this work was conducted.

ACKNOWLEDGEMENTS

This project was financially supported by the Chinese National Basic Research Priorities Program (No. 2005CB523402) and a key grant from the National Natural Science Foundation of China (No. 90209005). The authors thank Mr. Egon Willighagen for critical review of the second draft of the manuscript.

Table 9: Summary of Predicted Values by Hou's, Ponce's and SVM model

Compound	Caco-2[a]	Prediction			
		MLR[b]	SVM[c]	Hou's[d]	Ponce's[e]
bosentan	-5.98	-5.43	-5.29		
ceftriaxone	-6.88	-5.32	-5.48		
coumarin	-4.11	-4.44	-4.54		
amoxicillin	-6.1	-5.65	-5.84	-6.16	
antipyrine	-4.55	-4.45	-4.51	-4.82	
cyclosporine	-6.05	-5.82	-5.39	-5.81	
enalapril	-5.64	-5.04	-4.89	-5.66	
epinephrine	-6.02	-5.56	-5.72	-5.47	
diltiazem	-4.38	-4.57	-4.87	-4.84	-3.17
fleroxacin	-4.81	-4.76	-4.87	-5.39	-3.95
furosemide	-6.5	-5.81	-5.67	-5.81	-8.74
guanabenz	-4.5	-5.24	-5.35	-4.63	-6.68
guanoxan	-4.71	-5.27	-5.41	-5.39	-6.69
lidocaine	-4.21	-4.75	-4.67	-4.45	-4.83
mibefradil	-4.87	-4.97	-4.74	-5.06	-4.83
nitrendipine	-4.77	-4.73	-4.93	-5.08	-4.87
proscillaridin	-6.2	-5.77	-5.27	-5.42	-5.63
remikiren	-6.13	-6.17	-5.53	-5.36	-8.33
saquinavir	-6.26	-5.87	-5.55	-5.39	-9.32
sulpiride	-6.16	-5.47	-5.13	-5.81	-7.76
theophylline	-4.35	-4.66	-4.74	-5.06	-4.65
tiacrilast	-4.9	-4.97	-4.9	-5.68	-3.89
verapamil	-4.58	-4.52	-4.84	-4.87	-3.17
r (Ponce's Set)		0.79	0.84	0.74	0.78
UME(Ponce's Set)		0.46	0.44	0.52	1.29
r (Hou's Set)		0.79	0.83	0.78	
UME(Hou's Set)		0.43	0.42	0.45	
r (whole set)		0.76	0.85		
UME (whole set)		0.49	0.46		

[a] Caco-2, experimental Caco-2 apparent permeability (cm/s), $logP_{app}$; [b]Prediction by the MLR model described in this paper. [c] Prediction by the SVM model described in this paper. [d] Prediction by Hou's model. [e] Prediction by Ponce's model. *UME* is unsigned mean error.

REFERENCES

[1] F. Lombardo, E. Gifford, and M. Y. Shalaeva. In silico ADME prediction: data, models, facts andmyths. *Mini Rev Med Chem* **3**: 861-75 (2003).

[2] S. Ekins, Y. Nikolsky, and T. Nikolskaya. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol Sci* **26**: 202-9 (2005).

[3] F. Yamashitaand M. Hashida. In silico approaches for predicting ADME properties of drugs. *Drug Metab Pharmacokinet* **19**: 327-38 (2004).

[4] A. R. Hilgers, D. P. Smith, J. J. Biermacher, J. S. Day, J. L. Jensen, S. M. Sims, W. J. Adams, J. M. Friis, J. Palandra, J. D. Hosley, E. M. Shobe, and P. S. Burton. Predicting oral absorption of drugs: a case study with a novel class of antimicrobial agents. *Pharm Res* **20**: 1149-55 (2003).

[5] Y. H. Zhao, M. H. Abraham, J. Le, A. Hersey, C. N. Luscombe, G. Beck, B. Sherborne, and I. Cooper. Rate-limited steps of human oral absorption and QSAR studies. *Pharm Res* **19**: 1446-57 (2002).

[6] K. Obata, K. Sugano, R. Saitoh, A. Higashida, Y. Nabuchi, M. Machida, and Y. Aso. Prediction of oral drug absorption in humans by theoretical passive absorption model. *Int J Pharm* **293**: 183-92 (2005).

[7] J. P. Bai, A. Utis, G. Crippen, H. D. He, V. Fischer, R. Tullman, H. Q. Yin, C. P. Hsu, L. Jiang, and K. K. Hwang. Use of classification regression tree in predicting oral absorption in humans. *J Chem Inf Comput Sci* **44**: 2061-9 (2004).

[8] J. V. Turner, D. J. Maddalena, and S. Agatonovic-Kustrin. Bioavailability prediction based on molecular structure for a diverse series of drugs. *Pharm Res* **21**: 68-82 (2004).

[9] Y. C. Martin. A bioavailability score. *J Med Chem* **48**: 3164-70 (2005).

[10] N. P. Vermeulen. Prediction of drug metabolism: the case of cytochrome P450 2D6. *Curr Top Med Chem* **3**: 1227-39 (2003).

[11] G. Cruciani, E. Carosati, B. De Boeck, K. Ethirajulu, C. Mackie, T. Howe, and R. Vianello. MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J Med Chem* **48**: 6970-9 (2005).

[12] M. Shen, Y. Xiao, A. Golbraikh, V. K. Gombar, and A. Tropsha. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J Med Chem* **46**: 3013-20 (2003).

[13] Y. H. Wang, Y. Li, S. L. Yang, and L. Yang. Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. *J Chem Inf Model* **45**: 750-7 (2005).

[14] H. Bohets, P. Annaert, G. Mannens, L. Van Beijsterveldt, K. Anciaux, P. Verboven, W. Meuldermans, and K. Lavrijsen. Strategies for absorption screening in drug discovery and development. *Curr Top Med Chem* **1**: 367-83 (2001).

[15] P. Stenberg, U. Norinder, K. Luthman, and P. Artursson. Experimental and computational screening models for the prediction of intestinal drug absorption. *J Med Chem* **44**: 1927-37 (2001).

[16] K. Palm, K. Luthman, A. L. Ungell, G. Strandlund, and P. Artursson. Correlation of drug absorption with molecular surface properties. *J Pharm Sci* **85**: 32-9 (1996).

[17] U. Norinder, T. Osterberg, and P. Artursson. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharm Res* **14**: 1786-91 (1997).

[18] G. Camenisch, J. Alsenz, H. van de Waterbeemd, and G. Folkers. Estimation of permeability by passive diffusion through Caco-2 cell monolayers using the drugs' lipophilicity and molecular weight. *Eur J Pharm Sci* **6**: 317-24 (1998).

[19] S. D. Pickett, I. M. McLay, and D. E. Clark. Enhancing the hit-to-lead properties of lead optimization libraries. *J Chem Inf Comput Sci* **40**: 263-72 (2000).

[20] A. Kulkarni, Y. Han, and A. J. Hopfinger. Predicting Caco-2 cell permeation coefficients of organic molecules using membrane-interaction QSAR analysis. *J Chem Inf Comput Sci* **42**: 331-42 (2002).

[21] S. Fujiwara, F. Yamashita, and M. Hashida. Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network. *Int J Pharm* **237**: 95-105 (2002).

[22] Ponce Y.M., Perez M.A.C., Zaldivar V.R., Ofori E., and M. L.A. Total and Local Quadratic Indices of the "Molecular Pseudograph's Atom Adjacency Matrix". Application to Prediction of Caco-2

Permeability of Drugs. *Int. J. Mol. Sci* 512-536 (2003).

[23] T. J. Hou, W. Zhang, K. Xia, X. B. Qiao, and X. J. Xu. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J Chem Inf Comput Sci* **44**: 1585-600 (2004).

[24] Y. Marrero Ponce, M. A. Cabrera Perez, V. Romero Zaldivar, H. Gonzalez Diaz, and F. Torrens. A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. *J Pharm Pharm Sci* **7**: 186-99 (2004).

[25] H. H. Refsgaard, B. F. Jensen, P. B. Brockhoff, S. B. Padkjaer, M. Guldbrandt, and M. S. Christensen. In silico prediction of membrane permeability from calculated molecular parameters. *J Med Chem* **48**: 805-11 (2005).

[26] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* **43**: 493-500 (2003).

[27] R. Guha. Using the CDK as a backend to R. *CDK news* **2**: 2-6 (2005).

[28] C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. Willighagen. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics *(In Press)*. *Curr Pharm Des* **12**: (2006).

[29] S. Bandyopadhyay, A. Bagchi, and U. Maulik. Active site driven ligand design: an evolutionary approach. *J Bioinform Comput Biol* **3**: 1053-70 (2005).

[30] V. J. Gillet, P. Willett, and J. Bradshaw. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J Chem Inf Comput Sci* **38**: 165-79 (1998).

[31] R. D. Brownand Y. C. Martin. Designing combinatorial library mixtures using a genetic algorithm. *J Med Chem* **40**: 2304-13 (1997).

[32] S. Sun, P. D. Thomas, and K. A. Dill. A simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng* **8**: 769-78 (1995).

[33] C. M. Oshiro, I. D. Kuntz, and J. S. Dixon. Flexible ligand docking using a genetic algorithm. *J Comput Aided Mol Des* **9**: 113-30 (1995).

[34] R. L. Hauptand S. E. Haupt. *Practical Genetic Algorithms (2nd ed)*, John Wiley & Sons, Inc., Hoboken, 2004.

[35] S. J. Jung, S. O. Choi, S. Y. Um, J. I. Kim, H. Y. Choo, S. Y. Choi, and S. Y. Chung. Prediction of the permeability of drugs through study on quantitative structure-permeability relationship. *J Pharm Biomed Anal* (2006).

[36] V. Venkatraman, A. R. Dalby, and Z. R. Yang. Evaluation of mutual information and genetic programming for feature selection in QSAR. *J Chem Inf Comput Sci* **44**: 1686-92 (2004).

[37] B. Ustun, W. J. Melssen, M. Oudenhuijzen, and L. M. C. Buydens. Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta* **544**: 292-305 (2005).

[38] V. Vapnik. *Statistical Learning Theory*, Springer, N.Y., 1998.

[39] H. X. Liu, R. J. Hu, R. S. Zhang, X. J. Yao, M. C. Liu, Z. D. Hu, and B. T. Fan. The prediction of human oral absorption for diffusion rate-limited drugs based on heuristic method and support vector machine. *J Comput Aided Mol Des* **19**: 33-46 (2005).

[40] P. Lindand T. Maltseva. Support vector machines for the estimation of aqueous solubility. *J Chem Inf Comput Sci* **43**: 1855-9 (2003).

[41] Y. Xue, C. W. Yap, L. Z. Sun, Z. W. Cao, J. F. Wang, and Y. Z. Chen. Prediction of P-glycoprotein substrates by a support vector machine approach. *J Chem Inf Comput Sci* **44**: 1497-505 (2004).

[42] J. M. Kriegl, T. Arnhold, B. Beck, and T. Fox. A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J Comput Aided Mol Des* **19**: 189-201 (2005).