ELSEVIER

International Journal of Pharmaceutics 237 (2002) 95–105

international journal of pharmaceutics

www.elsevier.com/locate/ijpharm

Prediction of Caco-2 cell permeability using a combination of MO-calculation and neural network

Shin-ichi Fujiwara, Fumiyoshi Yamashita, Mitsuru Hashida *

Department of Drug Delivery Research, Graduate School of Pharmaceutical Sciences, Kyoto University, Yoshidashimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan

Received 4 October 2001; received in revised form 27 December 2001; accepted 22 January 2002

Abstract

In the present study, we developed an approach involving a combination of molecular orbital (MO) calculation and neural network to predict Caco-2 cell permeability (log P_{app}) from the molecular structure of compounds. For a total of 87 compounds with log P_{app} values obtained from the literature, three-dimensional molecular structures were determined by MO-calculation, and then five molecular descriptors were obtained, namely, the dipole moment, polarizability, sum of charges of nitrogen atoms (sum(N)), oxygen atoms (sum(O)), and hydrogen atoms bonding to nitrogen or oxygen atoms (sum(H)). The correlation between these five molecular descriptors and log P_{app} was analyzed using a feed-forward back-propagation neural network with a configuration of 5-4-1 for input, hidden, and output layers found suitable for predicting Caco-2 cell permeability. A leave-one-out cross-validation procedure revealed that the neural network model possesses a fairly good predictability as far as Caco-2 cell permeability is concerned (predictive root mean square error (RMSE) = 0.507), and better than the simple and quadratic regression model (predictive RMSE = 0.584 and 0.568, respectively).

Keywords: Intestinal absorption; Caco-2 cells; Permeability; Molecular structure; Neural networks

1. Introduction

The oral route is generally preferred for drug administration because of its ease and good patient compliance. In a drug discovery setting, therefore, it is important to design and develop compounds that can be absorbed effectively through the intestinal epithelium. However, many of the compounds derived from combinatorial synthesis and high throughput screening have inappropriate properties for oral absorption, such as low solubility and low permeability, so that the rate of success in drug development is quite low (Lipinski et al., 1997). For efficient drug discovery and development, the design of suitable libraries prior to synthesis, i.e. the selection of subsets, is required. Computational methods that can predict oral absorption of novel compounds would greatly assist this process.

* Corresponding author. Tel.: +81-75-753-4525; fax: +81-75-753-4575.

E-mail address: hashidam@pharm.kyoto-u.ac.jp (M. Hashida).

0378-5173/02/$ - see front matter
PII: S0378-5173(02)00045-5

Wessel et al. (1998) recently reported a quantitative structure–property relationship (QSPR) model to predict human fractional absorption data for a large and diverse set of drugs and 'drug-like' compounds, where the error in prediction of human fractional absorption was 16.0% for the testing dataset. Hirono et al. (1994) reported quantitative property–bioavailability relationships for 188 different compounds, where individual equations for three categories of the data possess a reasonable accuracy. Generally speaking, however, prediction of human oral bioavailability is difficult, because it depends on several distinctive processes such as dissolution, membrane transport, and metabolism.

In vitro experimental methods make it possible to examine large numbers of samples, decrease the quantity of the compounds required, and give a clearer interpretation of results. Therefore, in vitro screening, e.g. Caco-2 cell permeability measurements, is routinely performed in an early stage of drug discovery (Hildalgo et al., 1989; Artursson et al., 1996; Delie and Rubas, 1997). To develop specialized expert systems to predict these properties would greatly help the design of compound libraries.

Several investigators have explored QSPR involving Caco-2 cell permeability. In their studies, various types of molecular descriptors have been introduced to the QSPR modeling, including size and hydrogen-bonding descriptors (Waterbeemd and Camenisch, 1996), (dynamic) polar surface area (PSA) (Waterbeemd and Kansy, 1992; Palm et al., 1996; Krarup et al., 1998), and Molsurf-derived descriptors (Norinder et al., 1997). These QSPR models can predict Caco-2 cell permeability with a reasonable accuracy, although the number of compounds in the datasets is limited.

In this study, we tried to develop a QSPR model for a larger set of Caco-2 cell permeability data obtained from different sources. We calculated molecular descriptors of structurally diverse compounds by a semi-empirical molecular orbital (MO)-calculation method, and then applied an artificial neural network to the multivariate analysis between molecular descriptors and Caco-2 cell permeability. An artificial neural network is a computer-based system derived from a simplified concept of the brain, that have been proven to be effective in predicting the octanol/water partition coefficient (Breindl et al., 1997), oral bioavailability (Wessel et al., 1998), and clinical pharmacokinetic data (Brier et al., 1995) for a number of drugs, as well as the influence of pharmaceutical formulations (Kesavan and Peck, 1996). We compared the neural network approach and multiple linear regression with respect to the predictability of Caco-2 cell permeability.

2. Materials and methods

2.1. Molecular descriptors for QSPR analysis

The structures of the compounds were built with Chem 3D Pro Ver. 5.0 software (CambridgeSoft Co., Cambridge, MA) and modeled in their neutral forms. Geometry optimization was performed initially by molecular mechanics (MM2) force field and subsequently using the AM1 Hamiltonian of a semi-empirical MO PACkage (MOPAC97). Molecular descriptors of each compound in its optimum geometry were then calculated. These descriptors included dipole moment, polarizability, sum of charges of nitrogen atoms (sum(N)), oxygen atoms (sum(O)), and hydrogen atoms bonding to nitrogen or oxygen atoms (sum(H)).

2.2. Prediction models generated by neural network

The calculated molecular descriptors and $\log P_{app}$ were correlated using a feed-forward 3-layered neural network. The input layer consisted of five molecular descriptor variables (dipole moment, polarizability, sum(N), sum(O), and sum(H)), whereas the output layer consisted of the response variable, the logarithm of the Caco-2 cell permeability coefficient ($\log P_{app}$). Before training was started, the input values were scaled between -1 and 1, while the output data were scaled between 0 and 1. An error back-propagation algorithm was used for network training, where the learning rate and momentum parameters were 0.05 and 0.7, respectively. The range of

weights was varied from − 3 to 3. The goodness-of-fit was evaluated by the predictive root mean square error (RMSE) defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\log P_{\mathrm{app,observed}} - \log P_{\mathrm{app,calculated}})^2}{n}}$$

The predictability of a neural network model was evaluated using a leave-one-out procedure. This method systematically removed one data point at a time from the data set. A neural network model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data point. This procedure was repeated until a complete set of predicted values was obtained.

2.3. Analysis of variance for data compilation problem

To check whether each sub-dataset from different sources could be combined with the others, analysis of variance for the residual sum of squares (RSE) was carried out. First, for an entire dataset, the RSE was calculated using a neural network model. Next, a set of data from a source was taken out, and then the neural network model was re-optimized followed by calculation of the RSE. The difference in the RSE between before and after taking out the sub-dataset (ΔRSE) was analyzed by an analysis of variance. The F value can be calculated using the following equation:

$$F_{m,l-n} = \frac{\Delta\mathrm{RSE}/m}{\mathrm{RSE}/(l-n)}$$

where l and m are the numbers of data in the entire dataset and the removed sub-dataset, respectively; n is the degree-of-freedom of a neural network model. If the probability corresponding to the F value was less than 0.05, the sub-dataset was regarded significantly different from the rest datasets.

3. Results

Table 1 summarizes the Caco-2 cell permeability data ($\log P_{\mathrm{app}}$) for 87 structurally diverse compounds, obtained from nine different sources (Artursson, 1990; Artursson and Karlsson, 1991; Haeberlin et al., 1993; Rubas et al., 1993; Hovgaard et al., 1995; Augustijns et al., 1996; Collett et al., 1996; Yee, 1997; Yazdanian et al., 1998). Even when $\log P_{\mathrm{app}}$ for the compounds obtained from two or more sources, the $\log P_{\mathrm{app}}$ values were not averaged, so that distinctiveness of each sub-dataset could statistically be evaluated. Therefore, 129 $\log P_{\mathrm{app}}$ values were used for the following multivariate analyses.

Table 2 summarizes five molecular descriptors calculated by the MO method for the compounds used in this study. This dataset included compounds that are very different in structure and membrane permeability. Each parameter of the compounds exhibited a wide range: dipole moment (μ, 0.62–11.05), polarizability (α, 10.9–335.2), sum(N) (− 1.973–0), sum(O) (− 3.861–0), sum(H) (0–1.672) and $\log P_{\mathrm{app}}$ (− 6.96– − 3.88). Table 3 summarizes the correlation coefficient (r) between any two molecular descriptors. The average was 0.324, and the highest r value was 0.671. Thus, severe multicollinearity of the descriptors was not observed.

Multiple linear regression analysis of 129 $\log P_{\mathrm{app}}$ values using five molecular descriptors as explanatory variables gave the following equation:

$$\begin{aligned}\log P_{\mathrm{app}} = {} & (-4.17 \pm 0.33) + (-0.104 \pm 0.055)\mu \\ & + (1.40 \pm 2.14) \times 10^{-3}\alpha \\ & + (0.227 \pm 0.231)\mathrm{sum(N)} \\ & + (-0.020 \pm 0.163)\mathrm{sum(O)} \\ & + (-1.297 \pm 0.360)\mathrm{sum(H)} \\ & (n = 129, R^2 = 0.559, s = 0.562, F_{5,123} \\ & = 31.24, P < 1\mathrm{E}-19) \end{aligned} \quad (1)$$

where the values in parentheses are regression coefficients with 95% confidence intervals; n is the number of drugs; R is the correlation coefficient; and s is the standard error. To consider combined effects of the descriptors, quadratic and interactive terms were also incorporated in the following regression analysis. The regression model optimized by forward–backward stepwise selection was as follows:

Table 1
Caco-2 cell permeability (log P_{app}) for 87 structurally diverse compounds obtained from nine sources

Compound	log P_{app} (cm/s)[a]								
	A	B	C	D	E	F	G	H	I
Acebutolol					−5.35				−6.29
Acebutolol ester[b]					−4.11				
Acetylsalicylic acid	−5.62							−4.51	−5.04
Acyclovir								−5.70	−6.60
Alprenolol	−4.39	−4.39			−4.12				−4.60
Alprenolol ester[b]					−3.97				
Aminopyrine									−4.44
Artemisinin						−4.52			
Artesunate						−5.40			
Atenolol	−6.70	−6.69					−6.33		−6.28
Azithromycin								−5.98	
Benzyl penicillin								−5.71	
Betaxolol					−4.31				
Betaxolol ester[b]					−4.02				
Bremazocine									−5.10
Caffeine								−4.30	−4.51
Chloramphenicol								−4.69	
Chlorothiazide									−6.72
Chlorpromazine									−4.70
Cimetidine							−6.30	−5.51	−5.86
Clonidine								−4.52	−4.66
Corticosterone	−4.26								−4.67
Desipramine								−4.67	−4.61
Dexamethasone	−4.90		−4.57					−4.63	−4.91
Dexamethasone-β-D-glucoside			−6.36						
Dexamethasone-β-D-glucuronide			−5.94						
Diazepam								−4.15	−4.48
Dopamine									−5.03
Doxorubicin								−6.80	
Erythromycin								−5.43	
Estradiol									−4.77
Felodipine	−4.64								
Fluconazole								−4.53	
Ganciclovir				−5.57					−6.42
Griseofulvin									−4.44
H216/44		−6.04							
Hydrochlorothiazide							−5.82		−6.29
Hydrocortisone	−4.67			−4.45					−4.85
Ibuprophen								−4.28	
Imipramine								−4.85	
Indomethacin									−4.69
Labetalol									−5.03
Mannitol	−6.75			−5.49				−6.19	−6.42
Meloxicam									−4.71
Methanol				−3.88					
Methotrexate								−5.92	
Methylscopolamine									−6.16
Metoprolol	−4.57	−4.57							−4.63

Table 1 (*Continued*)

Compound	log P_{app} (cm/s)[a]								
	A	B	C	D	E	F	G	H	I
Nadolol									−5.41
Naloxone								−4.55	
Naproxen				−4.13					
Nevirapine									−4.52
Nicotine									−4.71
Olsalazine	−6.96								
Oxprenolol					−4.18				
Oxprenolol ester[b]					−4.01				
Phencyclidine									−4.61
Phenytoin									−4.57
Pindolol									−4.78
Pirenzepine									−6.36
Piroxicam									−4.45
Practolol	−6.05	−6.05							
Prazocin								−4.36	
Progesterone				−4.10					−4.63
Propranolol	−4.38	−4.38			−4.08			−4.56	−4.66
Propranolol ester[b]					−3.98				
Quinidine								−4.69	
Ranitidine									−6.31
Salicylic acid	−4.92								−4.66
Scopolamine									−4.93
Sucrose									−5.77
Sulphasalazine	−6.89								−6.52
Sumatriptan								−5.52	
Taurocholic acid								−4.46	
Telmisartan									−4.82
Tenidap								−4.29	
Terbutaline	−6.42								−6.33
Testosterone	−4.29							−4.14	−4.60
Timolol					−4.35				−4.89
Timolol ester[b]					−4.10				
Trovafloxacin								−4.52	
Uracil									−5.37
Urea									−5.34
Valproic acid								−4.32	
Warfarin	−4.42								−4.68
Zidovudine									−5.16
Ziprasidone								−4.91	

[a] A: Artursson (1990); B: Artursson and Karlsson (1991); C: Haeberlin et al. (1993); D: Rubas et al. (1993); E: Hovgaard et al. (1995); F: Augustijns et al. (1996); G: Collett et al. (1996); H: Yee (1997); I: Yazdanian et al. (1998).

[b] *O*-cyclopropane carboxylic acid ester.

$$\begin{aligned}\log P_{app} = {} & (-3.83 \pm 0.24) \\ & + (-1.660 \pm 0.402)\mathrm{sum(H)} \\ & + (-0.177 \pm 0.067)\mu \\ & + (-6.07 \pm 3.61) \times 10^{-2}\mu\mathrm{sum(O)} \\ & + (-0.319 \pm 0.207)\mathrm{sum(N)sum(O)} \\ & + (-0.471 \pm 0.486)\mathrm{sum(N)sum(H)} \\ & (n = 129, R^2 = 0.582, s = 0.547,\ F_{5,123} \\ & = 34.18, P < 1\mathrm{E}-21) \end{aligned} \quad (2)$$

Table 2
Computed molecular descriptors for the compounds used for the present analysis

Compound	μ[a]	α[b]	sum(N)[c]	sum(O)[d]	sum(H)[e]
Acebutolol	6.28	173.9	−0.662	−1.132	0.590
Acebutolol ester[f]	3.48	205.1	−0.667	−1.383	0.377
Acetylsalicylic acid	1.45	85.1	0.000	−1.211	0.241
Acyclovir	5.96	117.7	−1.219	−0.900	0.857
Alprenolol	1.99	131.1	−0.301	−0.569	0.372
Alprenolol ester[f]	2.96	165.1	−0.306	−0.789	0.156
Aminopyrine	4.38	133.6	−0.514	−0.304	0.000
Artemisinin	6.14	124.3	0.000	−1.043	0.000
Artesunate	4.19	157.7	0.000	−1.759	0.242
Atenolol	4.13	133.9	−0.752	−0.883	0.795
Azithromycin	4.86	330.0	−0.581	−3.519	1.057
Benzyl penicillin	4.51	157.2	−0.678	−1.208	0.481
Betaxolol	1.14	190.8	−0.297	−0.843	0.371
Betaxolol ester[f]	2.46	157.0	−0.301	−1.136	0.164
Bremazocine	2.43	162.3	−0.196	−0.516	0.414
Caffeine	3.78	100.7	−0.880	−0.670	0.000
Chloramphenicol	6.00	132.9	−0.378	−0.999	0.693
Chlorothiazide	4.71	140.5	−1.872	−3.444	0.674
Chlorpromazine	2.67	180.2	−0.492	0.000	0.000
Cimetidine	9.74	141.3	−1.336	0.000	0.671
Clonidine	0.62	111.4	−0.741	0.000	0.399
Corticosterone	1.48	161.5	0.000	−1.191	0.412
Desipramine	2.31	157.6	−0.548	0.000	0.148
Dexamethasone	1.96	175.2	0.000	−1.490	0.625
Dexamethasone-β-D-glucoside	5.01	236.2	0.000	−3.028	1.294
Dexamethasone-β-D-glucuronide	3.89	240.2	0.000	−3.305	1.367
Diazepam	3.05	161.3	−0.468	−0.310	0.000
Dopamine	2.16	78.9	−0.332	−0.452	0.698
Doxorubicin	4.02	259.5	−0.337	−2.846	1.441
Erythromycin	9.47	314.3	−0.243	−3.861	1.063
Estradiol	2.67	139.1	0.000	−0.564	0.411
Felodipine	4.73	179.7	−0.238	−1.203	0.208
Fluconazole	1.95	138.6	−0.960	−0.329	0.218
Ganciclovir	4.59	128.7	−1.206	−1.260	1.091
Griseofulvin	3.16	172.3	0.000	−1.297	0.000
H216/44	3.41	237.3	−0.990	−1.804	0.784
Hydrochlorothiazide	9.15	136.8	−1.922	−3.440	0.884
Hydrocortisone	2.85	164.9	0.000	−1.509	0.619
Ibuprophen	1.84	104.0	0.000	−0.677	0.241
Imipramine	1.25	164.6	−0.503	0.000	0.000
Indomethacin	1.45	196.0	−0.220	−1.176	0.241
Labetalol	5.03	176.0	−0.750	−0.982	1.094
Mannitol	4.48	62.8	0.000	−1.936	1.266
Meloxicam	4.58	199.4	−1.129	−2.334	0.551
Methanol	1.59	10.9	0.000	−0.321	0.194
Methotrexate	5.35	264.7	−1.973	−1.676	1.589
Methylscopolamine	11.05	171.6	−0.013	−1.077	0.209
Metoprolol	0.64	136.4	−0.294	−0.826	0.368
Nadolol	3.25	148.4	−0.292	−1.174	0.767
Naloxone	4.78	162.7	−0.270	−0.959	0.415
Naproxen	1.25	130.4	0.000	−0.885	0.241
Nevirapine	2.60	157.2	−0.846	−0.344	0.244

Table 2 (*Continued*)

Compound	μ[a]	α[b]	sum(N)[c]	sum(O)[d]	sum(H)[e]
Nicotine	3.02	86.0	−0.403	0.000	0.000
Olsalazine	1.93	177.8	−0.139	−1.885	1.032
Oxprenolol	2.13	161.6	−0.302	−0.843	0.394
Oxprenolol ester[f]	0.99	129.5	−0.302	−1.068	0.155
Phencyclidine	0.93	131.6	−0.300	0.000	0.000
Phenytoin	3.42	139.0	−0.817	−0.617	0.526
Pindolol	1.60	132.6	−0.519	−0.543	0.610
Pirenzepine	5.62	195.1	−1.284	−0.590	0.244
Piroxicam	4.03	190.0	−1.161	−2.348	0.545
Practolol	4.49	140.4	−0.651	−0.878	0.586
Prazocin	3.36	226.9	−1.404	−0.736	0.398
Progesterone	2.67	148.2	0.000	−0.573	0.000
Propranolol	2.04	144.3	−0.302	−0.569	0.371
Propranolol ester[f]	3.72	177.3	−0.300	−0.831	0.155
Quinidine	1.96	180.6	−0.348	−0.528	0.207
Ranitidine	5.61	156.6	−0.840	−0.078	0.392
Salicylic acid	1.17	67.5	0.000	−0.955	0.510
Scopolamine	1.25	140.6	−0.253	−1.174	0.207
Sucrose	3.80	136.0	0.000	−3.323	1.672
Sulphasalazine	5.64	240.9	−1.059	−2.658	0.765
Sumatriptan	3.11	160.7	−1.260	−1.732	0.448
Taurocholic acid	5.84	223.9	−0.402	−3.794	1.082
Telmisartan	6.25	335.2	−0.741	−0.660	0.245
Tenidap	4.67	146.3	−0.341	−0.842	0.000
Terbutaline	3.96	109.9	−0.313	−0.812	0.804
Testosterone	4.14	134.3	0.000	−0.606	0.197
Timolol	1.53	153.4	−1.139	−0.783	0.361
Timolol ester[f]	4.45	186.4	−1.152	−0.998	0.156
Trovafloxacin	6.73	223.5	−0.979	−0.945	0.539
Uracil	4.19	54.6	−0.730	−0.664	0.532
Urea	3.18	24.2	−0.782	−0.371	0.810
Valproic acid	1.80	76.2	0.000	−0.681	0.239
Warfarin	1.87	167.9	0.000	−1.012	0.240
Zidovudine	2.05	130.9	−1.032	−1.298	0.475
Ziprasidone	2.80	235.6	−1.145	−0.308	0.257

[a] Dipole moment (μ).
[b] Polarizability (α).
[c] Sum of charges of nitrogen atoms (sum(N)).
[d] Oxygen atoms (sum(O)).
[e] Hydrogen atoms bonding to nitrogen or oxygen atoms (sum(H)).
[f] *O*-cyclopropane carboxylic acid ester.

Consideration of the quadratic and interactive terms improved the multiple correlation coefficient and standard error of the regression.

Utilization of a neural network model is another way to find a nonlinear relationship between causal factors and their results. In applying a neural network to prediction of $\log P_{app}$ from molecular descriptors, the number of units in the hidden layer needs to be optimized. Predictive RMSE obtained by a leave-one-out cross-validation procedure was used as a measure of the goodness of a neural network model. Table 4 summarizes RMSE to the entire dataset, predictive RMSE from leave-one-out cross-validation, and degree-of-freedom of 3-layer neural networks with varying number of units in a hidden layer.

The predictive RMSE was lowest for a neural network model with a 5-5-1 configuration; however, the value (0.500) was almost the same as that of a 5-4-1 neural network (0.507). Considering the degree-of-freedom of these two models, a 5-4-1 neural network was adopted as the optimal.

To compare the predictability between linear regression and neural network analyses, leave-one-out cross-validation was also carried out for the above-mentioned simple and quadratic regression models. Fig. 1 shows the relationship between observed and predicted log P_{app}. Predictive RMSEs were 0.584 and 0.568 for the simple and quadratic models, respectively, which was much larger than that of a 5-4-1 neural network (0.507). Thus, the neural network model gave a better predictability than the regression models.

Inter-laboratorical variability of Caco-2 cell permeability, which has been pointed out by several investigators (Artursson et al., 1996; Delie and Rubas, 1997), needs to be careful in compiling log P_{app} from a variety of sources. Therefore, an analysis of variance was carried out to check whether each sub-dataset could be combined with the others. As shown in Table 5, none of the sub-datasets were significantly different from the rest of an entire dataset. Although Caco-2 cell permeability had been measured under different conditions (Artursson, 1990; Artursson and Karlsson, 1991; Haeberlin et al., 1993; Rubas et al., 1993; Hovgaard et al., 1995; Augustijns et al., 1996; Collett et al., 1996; Yee, 1997; Yazdanian et al., 1998), any of the sub-datasets could not be excluded from a statistical point of view.

4. Discussion

Drug transport across cell membranes occurs primarily due to passive diffusion, where partitioning into and diffusion across the cell membranes are involved. Solute partitioning from an aqueous solution is determined by solute–solvent interactions that involve dipole–dipole interactions (Kamlet et al., 1987; Martin, 1993). The permanent dipole moment of a molecule is directly related to these interactions. The polarizability of the molecule is also responsible for solubility phenomena, since it determines inducible dipole moments. Hydrogen bonding is another important force to determine solute–solvent interactions. Several investigators (Kim et

Table 3
The correlation matrix between the molecular descriptors for the compounds

	μ	α	sum(N)	sum(O)	sum(H)
μ	1				
α	0.338	1			
sum(N)	−0.340	−0.200	1		
sum(O)	−0.374	−0.464	0.084	1	
sum(H)	0.325	0.263	−0.181	−0.671	1

Table 4
RMSE and degree-of-freedom for a feed-forward layered neural network with a 5-x-1 configuration ($x = 1, 2\ldots, 7$)

No. of units[a]	1	2	3	4	5	6	7
RMSE[b]	0.500	0.480	0.452	0.435	0.416	0.402	0.390
Predicted RMSE[c]	0.534	0.557	0.543	0.507	0.500	0.516	0.529
DF[d]	8	15	22	29	36	43	50

[a] Number of units in hidden layers.
[b] RMSE to an entire dataset ($n = 129$).
[c] Predictive RMSE obtained from leave-one-out cross-validation.
[d] Degree-of-freedom.

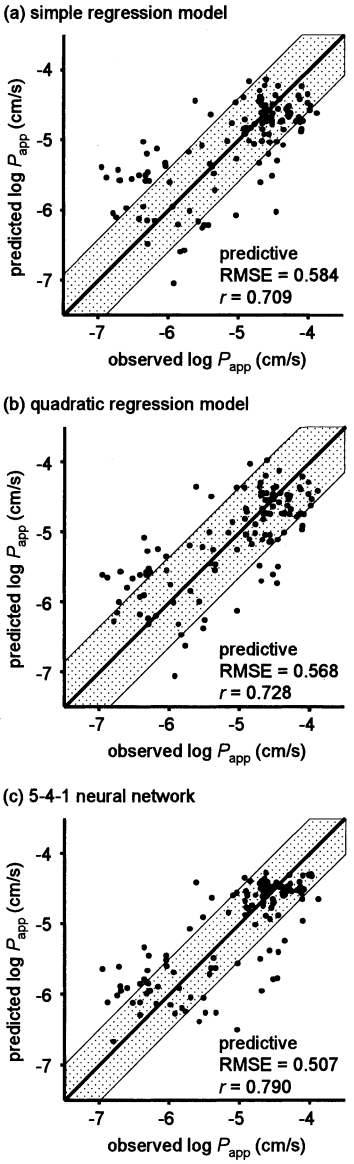

Fig. 1. Relationship between observed and predicted log P_{app} in the leave-one-out cross-validated prediction. (a)–(c) were obtained from analyses using a simple regression model, a quadratic regression model, and a 5-4-1 neural network, respectively. The shaded area expresses the range of the RMSE between experimental and predicted log P_{app}. Predictive RMSE and the correlation coefficient (r) between experimental and predicted log P_{app} are given in each graph.

al., 1996; Goodwin et al., 2001) have suggested that hydrogen-bonding potentials be a major factor in determining the Caco-2 cell permeability of solutes. This would be associated with the process of 'stripping' a molecule from its water of hydration. These properties of a molecule can be obtained from their 3-dimensional structure by MO-calculation. In this study, we tried to predict Caco-2 cell permeability from a physicochemical point of view, using MO-derived molecular descriptors (i.e. the dipole moment, polarizability, sum(N), sum(O) and sum(H)).

The molecular size of a solute is another important factor, determining the diffusivity in biological membranes. Therefore, the molecular volume or weight is often considered in the QSAR models for prediction of membrane permeability (Lien and Gao, 1995; Waterbeemd and Camenisch, 1996; Goodwin et al., 2001). It is known that the polarizability of a molecule is a good measure of its volume (Atkins, 1994). In fact, the calculated polarizability of the compounds currently investigated correlates closely with their molecular weight ($r^2 = 0.866$, the data not shown). Therefore, when the polarizability was used as a molecular descriptor, it would not be necessary to use the molecular volume (or weight) of the compound.

The relationship between causal factors and their results has often been analyzed by multiple linear regression analysis. However, if only the strict additivity of contributing descriptors were used, this would fail to take into account information involving the combined effect of several features. In addition, a nonlinear relationship might be needed when effects of heterogeneity of the permeation pathway are considered. This problem might be overcome to some degree by considering quadratic and interactive terms of the descriptors (Eq. (2)). Alternatively, a neural network can simulate a nonlinear relationship between causal factors and their results. When the leave-one-out cross-validation was carried out to check the predictability for Caco-2 cell permeability (Fig. 1), a neural network model gave a much better predictability than simple and quadratic regression models.

The PSA has been found to be a very effective descriptor for predicting Caco-2 cell permeability (Palm et al., 1996; Krarup et al., 1998) and human intestinal absorption (Palm et al., 1997; Clark, 1999; Pickett et al., 2000). It is interesting

Table 5
The analysis of variance for the RSE between before and after taking each sub-dataset using a 5-4-1 neural network model

Dataset	A	B	C	D	E	F	G	H	I
n[a]	17	6	3	3	12	2	6	29	51
ΔRSE[b]	5.61	1.01	0.39	0.52	3.05	0.60	2.34	6.72	12.40
F value[c]	1.352	0.687	0.535	0.704	1.040	1.238	1.594	0.949	0.996
P[d]	0.178	0.661	0.660	0.552	0.419	0.294	0.157	0.548	0.496

[a] The number of data.
[b] Difference in the RSE between before and after taking out the sub-dataset.
[c] F-test was carried out using RSE = 24.42 (degree-of-freedom = 100 (= 129 − 29)).
[d] Probability corresponding to the F value.

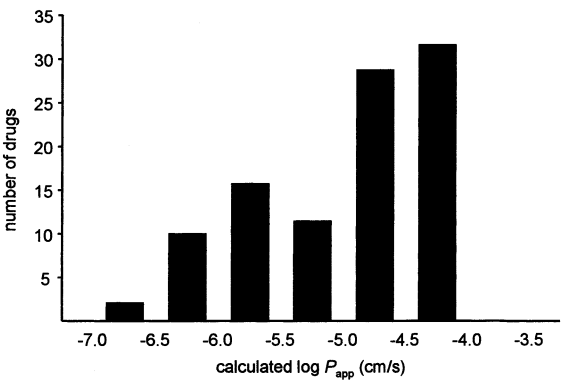

Fig. 2. Frequency distribution of log P_{app} calculated by the 5-4-1 neural network for 139 approved orally active drugs.

to compare the present approach with the PSA approach. Eighty-three log P_{app} data of the compounds, of which the PSA was available from the literature (Clark, 1999), were analyzed by these two approaches. When a leave-one-out cross-validation held for the PSA approach, the cross-validated predictive RMSE value was 0.622. Even when molecular weight was considered in addition to PSA (Waterbeemd and Camenisch, 1996), the predictability was not so much improved (predictive RMSE = 0.606). Our approach, that gave a predictive RMSE of 0.477, appears to be superior to these approaches.

To what extent do orally active drugs permeate Caco-2 cell monolayers? We estimated the Caco-2 cell permeability for 139 approved orally active drugs using our neural network model (Fig. 2). A distribution of the estimated log P_{app} for orally active drugs was negatively skewed (skewness of −0.790). These simulations suggest that most of the approved drugs easily cross the cell monolayers. This result would be reasonable, supposing that orally active drugs need to be effectively absorbed into the systemic circulation. In addition, approximately 90% of these compounds have a log P_{app} of more than −6.0. This could be a criterion for membrane permeability at an early stage of drug discovery.

In conclusion, the prediction of Caco-2 cell permeability from the molecular structure is possible using a combinatorial approach of MO-calculation and neural network. This approach would be useful in drug discovery settings where the likelihood of success in the later stages of drug development needs to be improved.

Acknowledgements

This research was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

References

Artursson, P., 1990. Epithelial transport of drugs in cell culture. I: a model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. J. Pharm. Sci. 79, 476–482.

Artursson, P., Karlsson, J., 1991. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. Biochem. Biophys. Res. Com. 175, 880–885.

Artursson, P., Palm, K., Luthman, K., 1996. Caco-2 monolayers in experimental and theoretical predictions of drug transport. Adv. Drug Deliv. Rev. 22, 67–84.

Atkins, P.W., 1994. Physical Chemistry, fifth ed. Oxford University Press, Oxford.

Augustijns, P., D'Hulst, A., Daele, J.V., Kinget, R., 1996. Transport of artemisinin and sodium artesnate in Caco-2 intestinal epithelial cells. J. Pharm. Sci. 85, 577–579.

Breindl, A., Beck, B., Clark, T., Glen, R.C., 1997. Prediction of the *n*-octanol/water coefficient, log *P*, using a combination of semiempirical MO-calculations and a neural network. J. Mol. Model 3, 142–155.

Brier, M.E., Zurada, J.M., Aronoff, G.R., 1995. Neural network predicted peak and trough gentamicin concentration. Pharm. Res. 12, 406–412.

Clark, D.E., 1999. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. J. Pharm. Sci. 88, 807–814.

Collett, A., Sims, E., Walker, D., He, Y.-L., Ayrton, J., Rowland, M., Warhurst, G., 1996. Comparison of HT29-18-C_1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. Pharm. Res. 13, 216–221.

Delie, F., Rubas, W.A., 1997. Human colonic cell line sharing similarities with enterocytes as a model to examine oral absorption. Crit. Rev. Ther. Drug Carrier Syst. 14, 221–286.

Goodwin, J.T., Conradi, R.A., Ho, N.F., Burton, P.S., 2001. Physicochemical determinants of passive membrane permeability: role of solute hydrogen-bonding potential and volume. J. Med. Chem. 44, 3721–3729.

Haeberlin, B., Rubas, W., Nolen, H.W. III, Friend, D.R., 1993. In vitro evaluation of dexamethasone-β-D-glucuronide for colon-specific drug delivery. Pharm. Res. 10, 1553–1562.

Hildalgo, I.J., Raub, T.J., Borchardt, R.T., 1989. Characterization of the human colon carcinoma cell line (Caco-2) as a model system for intestinal epithelial permeability. Gastroenterology 96, 736–749.

Hirono, S., Nakagome, I., Hirano, H., Matsushita, Y., Yoshii, F., Moriguchi, I., 1994. Non-congeneric structure-pharmacokinetic property correlation studies using fuzzy adaptive least-squares: oral bioavailability. Biol. Pharm. Bull. 17, 306–309.

Hovgaard, L., Brøndsted, H., Buur, A., Bundgaard, H., 1995. Drug delivery studies in Caco-2 monolayers. Synthesis, hydrolysis, and transport of *O*-cyclopropane carboxylic acid ester prodrugs of various β-blocking agents. Pharm. Res. 12, 387–392.

Kamlet, M.J., Doherty, R.M., Fiserova-Bergerova, V., Carr, P.W., Abraham, M.H., Taft, R.W., 1987. Solubility properties in biological media 9: prediction of solubility and partition of organic nonelectrolytes in blood and tissues from solvatochromic parameters. J. Pharm. Sci. 76, 14–17.

Kesavan, J.G., Peck, G.E., 1996. Pharmaceutical granulation and tablet formulation using neural networks. Pharm. Dev. Technol. 1, 391–404.

Kim, D.C., Burton, P.S., Borchardt, R.T., 1996. A correlation between the permeability characteristics of a series of peptides using an in vitro cell culture model (Caco-2) and those using an in situ perfused rat ileum model of the intestinal mucosa. Pharm. Res. 10, 1710–1714.

Krarup, H., Christensen, I.T., Hovgaard, L., Frokjaer, S., 1998. Predicting drug absorption from molecular surface properties based on molecular dynamics simulations. Pharm. Res. 15, 972–978.

Lien, E.J., Gao, H., 1995. QSAR analysis of skin permeability of various drugs in man as compared to in vivo and in vitro studies in rodents. Pharm. Res. 12, 583–587.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Deliv. Rev. 23, 3–25.

Martin, A., 1993. Physical Pharmacy, fourth ed. Lea & Febiger, Philadelphia.

Norinder, U., Österberg, T., Artursson, P., 1997. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parameterization and PLS statistics. Pharm. Res. 14, 1786–1791.

Palm, K., Luthman, K., Ungell, A.-L., Strandlund, G., Artursson, P., 1996. Correlation of drug absorption with molecular surface properties. J. Pharm. Sci. 85, 32–39.

Palm, K., Stenberg, P., Luthman, K., Artursson, P., 1997. Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm. Res. 14, 568–571.

Pickett, S.D., McLay, I.M., Clark, D.E., 2000. Enhancing the hit-to-lead properties of lead optimization libraries. J. Chem. Inf. Comput. Sci. 40, 263–272.

Rubas, W., Jezyk, N., Grass, G.M., 1993. Comparison of the permeability characteristics of a human colonic epithelial (Caco-2) cell line to colon of rabbit, monkey, and dog intestine and human drug absorption. Pharm. Res. 10, 113–118.

Waterbeemd, H., Camenisch, G., 1996. Estimation of Caco-2 cell permeability using calculated molecular descriptors. Quant. Struct.-Act Relat. 15, 480–490.

Waterbeemd, H., Kansy, M., 1992. Hydrogen-bonding capacity and brain penetration. Chimia 46, 299–303.

Wessel, D., Jurs, P.C., Tolan, J.W., Muskal, S.M., 1998. Prediction of human intestinal absorption of drug compounds from molecular structure. J. Chem. Inf. Comput. Sci. 38, 726–735.

Yazdanian, M., Glynn, S.L., Wright, J.L., Hawi, A., 1998. A correlating partitioning and Caco-2 cell permeability of structurally diverse small molecular weight compounds. Pharm. Res. 15, 1490–1494.

Yee, S., 1997. In vitro permeability across Caco-2 cells (colonic) can predict in vivo (small intestinal) absorption in man—fact or myth. Pharm. Res. 14, 763–766.