A Quantitative Structure–Property Relationship for Predicting Drug Solubility in PEG 400/Water Cosolvent Systems

Erik Rytting,[1,3] Kimberley A. Lentz,[1,4] Xue-Qing Chen,[2] Feng Qian,[2] and Srini Venkatesh[1]

Received August 21, 2003; accepted October 22, 2003

Purpose. A quantitative structure–property relationship (QSPR) was developed to predict drug solubility in binary mixtures of polyethylene glycol (PEG) 400 and water. The ability of the QSPR model to predict solubility was assessed and compared to the classic log-linear cosolvency model.

Methods. The solubility of 122 drugs, ranging in log *P* from –2.4 to 7.5, was determined in 0%, 25%, 50%, and 75% PEG (v/v in water) by the shake-flask method. Solubility data from 84 drugs were fit by linear regression using the following molecular descriptors: molecular weight, volume, radius of gyration, density, number of rotatable bonds, hydrogen-bond donors, and hydrogen-bond acceptors. The multiple linear regression model was optimized by a genetic algorithm guided selection method. The remaining 38 compounds were used to test the predictability of the model.

Results. QSPR-based models developed at each volume fraction with the training set compounds showed a reasonable correlation coefficient (*r*) of ~0.9 and a root mean square (rms) error of <0.5 log unit. The model predicted solubility values of ~78% of the testing set compounds within 1 log unit. The log-linear model was as effective as the QSPR-based model in predicting the testing set solubilities; however, many drugs, as expected, showed significant deviation from log-linearity.

Conclusions. The QSPR model requires only the chemical structure of the drug and has utility for guiding vehicle identification for early preclinical *in vivo* studies, especially when compound availability is limited and experimental data such as aqueous solubility and melting point are unknown. When experimental data are available, the log-linear model was verified to be a useful predictive tool.

KEY WORDS: cosolvent; *in silico*; PEG 400; prediction; QSPR model; solubility.

INTRODUCTION

Computational methods for solubility and other physicochemical properties enable the drug discovery–development interface to become more efficient, as these methods often require little or no experimental input. Thus, predictive tools have the ability to accelerate drug candidate optimization and hence the overall drug screening process (1). An early assessment of *in vivo* pharmacokinetics is often critical to quickly make go/no-go decisions on new candidates. However, in early discovery, the amounts of compound available for identification of preclinical dosing vehicles are usually limited. Models that predict the solubility of compounds in pharmaceutically relevant solvents can assist the pharmaceutical scientist in the development of preclinical drug formulations, which allow for the determination of critical absorption distribution metabolism excretion (ADME) and toxicology data.

One of the properties most crucial to candidate screening is the solubility of the compound. When the aqueous solubility of a drug candidate is inadequate to permit solution formulations, cosolvents are often used to improve solubility (2). Cosolvents disrupt the strong interactions between water molecules, and therefore reduce the surface tension, solubility parameter, and polarity of the aqueous phase (3). Many drugs become more soluble as the cosolvent decreases the ability of water to "squeeze out" the nonpolar solutes (3).

Polyethylene glycol (PEG) 400 is commonly used as a cosolvent in the pharmaceutical industry. Sweetana and Akers estimate that 10% of FDA-approved parenteral products contain cosolvents (2). In addition, most *in vivo* formulations at the preclinical stage, for both oral and intravenous administration, are solutions. When drugs are insoluble in water, PEG is an excipient of choice based on its good solubilization properties and overall acceptability in terms of side-effect profile (4).

Although the literature presents several models for predicting solubilization by cosolvents (5), many require the collection of experimental data, a luxury often not afforded at this stage in drug discovery where compound is in short supply. In the presentation of an excess free energy approach to estimating solubility in mixed solvents, Williams and Amidon (6) questioned the utility of the extended form of the regular solution equation (7) because it requires the experimental determination of the heat of fusion, melting point, molar volume, and solubility parameters of both solute and solvent; so much lab work may thwart the goal to estimate solubility with the least number of experiments (6).

Yalkowsky, Flynn, and Amidon presented the classic log-linear relationship for drug solubility in binary aqueous systems (8). The accuracy of the log-linear model has been proven for many drugs and cosolvents (3,9–12), and the simplicity of the model complements its utility. Nevertheless, some experimental data must be determined prior to its application.

This work focused on the development of a predictive solubility model at three volume fractions of PEG 400: 25%, 50%, and 75%. Molecular descriptors such as molecular weight, volume, density, radius of gyration, number of rotatable bonds, hydrogen-bond donors, and hydrogen-bond acceptors were used such that the solubility predictions were made solely on the basis of the drug's chemical structure. This model is advantageous in that the data for all 122 drugs used to develop and test the model were generated by a uniform experimental procedure. Furthermore, the drugs represent a wide range of compounds, with log *P* values from –2.4 to 7.5 and molecular weights from 111 to 614 g/mol. Additionally, the predictive power of this model was compared to the more traditional approach of the log-linear model.

[1] Discovery Pharmaceutics, Preclinical Candidate Optimization, Bristol-Myers Squibb Pharmaceutical Research Institute, Wallingford, Connecticut 06492.

[2] Discovery Pharmaceutics, Preclinical Candidate Optimization, Bristol-Myers Squibb Pharmaceutical Research Institute, Lawrenceville, New Jersey 08543.

[3] Present address: Department of Pharmaceutical Chemistry, University of Kansas, Lawrence, Kansas 66047.

[4] To whom correspondence should be addressed. (e-mail: kimberley.lentz@bms.com)

MATERIALS AND METHODS

Materials

Polyethlylene glycol 400 was purchased from J. T. Baker (Phillipsburg, NJ, USA); methyl sulfoxide, methanol, and acetonitrile were obtained from EM Science (Gibbstown, NJ, USA); 0.1 N sodium hydroxide solution was acquired from VWR Scientific Products (West Chester, PA, USA). All drugs studied (listed in Table I) were purchased from Sigma-Aldrich, Inc. (St. Louis, MO, USA). Only the free forms of the compounds were obtained (i.e., no salts or other derivatives were involved in the current study).

Experimental Methods

The crystallinity of each drug was verified by either differential scanning calorimetry, X-ray diffractometry, and/or microscopy. For each drug, an amount sufficient to ensure saturation was mixed with 200 μl of 25%, 50%, and 75% (v/v) PEG in separate vials. The vials were vortexed, then shaken at room temperature (23 ± 2°C) for at least 24 h in a Burrell Scientific WristAction Shaker (Pittsburgh, PA, USA) in order to obtain equilibrium. Equilibrium was confirmed by "spot checking" various samples at 24 and 48 h. To prevent any possible photodegradation, all vials were protected from light. Formation of PEG solvates and hydrates upon equilibration of the sample was not examined in this study.

The vials were spun in an Eppendorf Centrifuge 5402 (Westbury, NY, USA) at 14,000 rpm ($g \times 100$) for 15 min to separate the saturated solution from the excess solid. The supernatant from each vial was filtered and/or diluted as necessary for quantitation purposes. The diluted samples, along with an appropriate standard curve, were analyzed on a Waters 2690 HPLC (Milford, MA, USA) with a Waters 996 Photodiode Array Detector.

Model Development

The predictive model was comprised of three quantitative structure–property relationships (QSPRs) to predict drug solubility in each of the volume fractions of PEG used in the experiments. The experimental solubility values from a training set of 84 drugs were fit by linear regression to specific molecular descriptors, calculated from the compound's structure. The calculations were performed on a Silicon Graphics O2 IRIX 6 workstation (SGI, Mountain View, CA, USA) and the descriptors calculated using Cerius^2 software (Molecular Simulations, Inc., San Diego, CA, USA) (13).

Thirty-eight of the 122 drugs were not included in the model fitting; rather, these drugs were reserved as a testing set to validate the model's performance. These compounds were selected by a combination of Cerius^2's cluster analysis and random number generation to achieve representation of all molecule types, so as not to bias the results.

The most important aspect of developing the model was the selection of appropriate molecular descriptors. Although hundreds of parameters are available for molecular modeling, it was necessary to choose descriptors chemically and physically relevant to solubility characterization (13,14). Certain chemical properties such as melting point, heat of fusion, and aqueous solubility might produce better modeling results, but these were ignored because they require experimental measurements, thus thwarting the objective to accelerate the process of screening drug candidates in the absence of sufficient drug substance to perform such experiments. Other descriptors, including log P and polarizability, were left out because they are calculated using group contribution methods that would introduce additional error.

A linear model was initially generated using the 84 compound training set. However, this single group model was not able to encompass the complexity of the data set, due largely to the diversity in chemical space. Table II shows that r was ≤ 0.67 for each of the PEG 400 volume fractions. Thus, descriptor selection alone did not adequately generate a reliable QSPR model.

In an effort to improve predictability, it was necessary to divide or "bin" the compounds of the training set into groups to optimize the model. This was accomplished through use of a genetic algorithm guided selection method developed by Cho and Hermsmeier (13,15). Briefly, this program categorized the training set compounds into multiple groups based on the similarities of the molecular descriptor values. The genetic algorithm guided selection requires a minimum compound to variable ratio of 5:1 in each group to get statistically sound models. It must be emphasized that this model does not group compounds based on chemotype; thus, structurally similar compounds could potentially fall into different groups if they possess very different molecular descriptor values. The 84 training set compounds were divided into two groups, and seven molecular descriptors were selected to regress the model. It was also possible to divide the training set into three groups and use five variables, but these attempts did not improve the model's performance and therefore such approaches were not pursued.

Combinations of molecular descriptors were analyzed to achieve the best fit. As the genetic algorithm guided selection method classified the training set compounds into two groups, a linear regression model was obtained for each group at each of the three volume fractions of PEG in the study. The regression model was a linear combination of the following seven molecular descriptors: molecular weight, volume, density, radius of gyration, number of hydrogen-bond donors, hydrogen-bond acceptors, and rotatable bonds. Because we do not expect linear returns to scale, a logarithmic transformation for solubility was carried out in order to maintain the linear relationship between solubility and the independent variables.

Model Validation

The predictive ability of the model was tested by comparing the predicted solubility values for the testing set compounds with the corresponding experimental values. The compounds in the testing set were not used for model development and therefore represented unknown compounds. The model predicted the solubility of each testing set compound as follows: The molecular descriptors of a particular testing set compound were compared to the descriptors of all 84 compounds in the training set. The training set compound that was most similar to the testing set compound determined which of the two groups the testing set compound belonged to. The solubility for the testing set compound was then predicted with the appropriate group model. Specifically, the similarity between compounds was determined by Euclidean distance, d_{ij}, as follows:

Table I. Drugs Studied, *in silico* Modeling Assignments, and Log-Linear σ Values

Compound name	Model set	MW	log *P**	σ
Acetazolamide	Test	222.24	−0.3	1.90
Adenine	Train	135.13	−0.1	1.28
Adenosine	Train	267.24	−1.3	0.85
Allopurinol	Test	136.11	−1.3	1.16
p-Aminobenzoic acid	Test	137.14	0.0	1.76
Aminopyrine	Train	231.30	0.8	0.70
5-Aminosalicylic acid	Test	153.14	0.5	0.74
p-Aminosalicylic acid	Train	153.14	0.3	2.72
Ampicillin	Train	349.40	1.4	−0.13
Aspirin	Train	180.16	1.2	1.99
Atropine	Train	289.37	1.5	1.48
Azathioprine	Train	277.26	0.9	2.16
Baclofen	Test	213.66	1.6	−0.50
Benzamide	Train	121.14	0.7	1.25
Benzocaine	Train	165.19	2.5	2.83
Benzoic acid	Train	122.12	1.9	2.65
Biphenyl	Test	154.21	4.0	5.16
Bumetanide	Train	364.42	2.8	4.00
Butamben	Train	193.25	3.6	4.28
Butylparaben	Train	194.23	3.5	4.35
Caffeine	Train	194.19	−0.1	−0.27
DL-Camphor	Test	152.24	2.1	1.40
Carbamazepine	Train	236.27	2.7	3.36
Cephradine	Train	349.40	1.0	−1.05
Chloramphenicol	Train	323.13	1.0	2.01
Chlorthalidone	Train	338.76	−0.7	3.65
Chlorzoxazone	Train	169.57	2.2	2.93
Cimetidine	Train	252.34	0.4	0.72
Clofazimine	Train	473.40	7.5	5.65
Corticosterone	Test	346.47	1.8	1.68
Cortisone	Train	360.45	1.2	1.74
Cytosine	Train	111.10	−1.7	0.21
Dapsone	Train	248.30	0.9	4.33
Deoxycorticosterone	Train	330.47	3.4	2.77
Dexamethasone	Train	392.47	2.1	3.00
Diatrizoic acid	Train	613.92	1.6	0.00
Diflunisal	Train	250.20	4.3	3.89
Diosgenin	Train	414.63	5.7	1.38
5,5-Diphenylhydantoin	Test	252.27	2.5	4.11
Disopyramide	Train	339.48	2.9	0.98
Diuron	Test	233.10	2.8	1.69
Equilin	Train	268.35	3.5	4.29
Estradiol-17-alpha	Train	272.39	4.1	4.96
Estriol	Train	288.39	2.9	3.76
Estrone	Test	270.37	3.7	3.99
Ethylparaben	Test	166.18	2.4	3.47
Ethynylestradiol-17-alpha	Test	296.41	4.5	4.97
Fenbufen	Train	254.28	3.0	3.21
Flufenamic acid	Train	281.23	5.6	4.02
5-Fluorocytosine	Test	129.09	−1.8	−0.79
5-Fluorouracil	Test	130.08	−0.8	−0.68
Flurbiprofen	Test	244.26	4.1	4.70
Folic acid	Train	441.40	−2.1	2.35
Glafenine	Train	372.81	3.9	3.88
Griseofulvin	Train	352.77	2.4	6.15
Guaifenesin	Train	198.22	0.6	1.47
Guanine	Train	151.13	−0.9	0.07
Haloperidol	Test	375.87	4.1	4.90
Hydrochlorothiazide	Train	297.73	−0.1	3.15
Hydrocortisone	Train	362.47	1.4	2.00
Hydroflumethiazide	Train	331.28	0.5	2.75
Hyoscyamine	Test	289.37	1.5	0.81
Ibuprofen	Train	206.28	3.7	4.33
Indapamide	Train	365.83	2.1	4.08
Indoprofen	Train	281.31	1.7	3.55
Iopanoic acid	Train	570.93	5.2	5.71
Ketoprofen	Train	254.28	2.8	4.20
Khellin	Test	260.25	1.7	1.58
Linuron	Test	249.10	3.2	3.07
Mefenamic acid	Train	241.29	5.3	4.33
Methocarbamol	Test	241.24	0.5	−1.00
Methylparaben	Train	152.15	1.9	2.84
Metronidazole	Train	171.16	0.0	0.13
Minoxidil	Train	209.25	−1.5	0.72
Nadolol	Train	309.40	1.3	0.04
Nalidixic acid	Train	232.24	0.2	0.85
Naphthalene	Train	128.17	3.4	3.94
2-Naphthol	Train	144.17	2.7	3.51
Naproxen	Train	230.26	3.0	4.30
Nitrofurantoin	Train	238.16	−0.5	1.99
Norethisterone	Train	298.42	3.4	3.30
Norfloxacin	Train	319.33	1.5	0.36
Paracetamol	Train	151.16	0.3	1.85
Perphenazine	Test	403.97	4.5	3.92
Phenacetin	Test	179.22	1.6	2.19
Phenolphthalein	Train	318.33	3.3	6.02
Phenylbutazone	Train	308.38	3.5	3.02
Praziquantel	Test	312.41	3.6	1.91
Prednisolone	Test	360.45	1.7	2.56
Primidone	Train	218.25	−1.0	2.71
Progesterone	Train	314.47	4.0	2.50
Propylparaben	Train	180.20	2.9	3.80
Pyrazinamide	Test	123.11	−0.4	0.06
Quinidine	Test	324.42	3.4	1.73
Quinine	Train	324.42	3.4	2.12
Salicylamide	Train	137.14	1.4	2.65
Salicylic acid	Test	138.12	2.1	2.91
Spironolactone	Train	416.57	3.2	2.97
Strychnine	Test	334.42	1.7	2.45
Sulfacetamide	Train	214.24	−0.9	2.42
Sulfadiazine	Train	250.27	−0.1	2.95
Sulfamerazine	Test	264.30	0.3	2.53
Sulfamethazine	Train	278.33	0.8	2.16
Sulfamethoxazole	Train	253.28	0.9	3.80
Sulfanilamide	Train	172.20	−0.7	1.19
Sulfathiazole	Train	255.31	0.3	2.60
Sulindac	Test	356.41	3.6	3.63
Sulpiride	Test	341.42	0.5	1.47
Tenoxicam	Test	337.37	−0.3	2.41
Terfenadine	Test	471.68	6.9	4.25
Tetraethylthiuram disulfide	Test	296.52	4.0	3.65
Theobromine	Train	180.17	−0.8	0.07
Theophylline	Test	180.17	0.1	−0.15
Thiamphenicol	Train	356.22	−0.3	1.62
Thymine	Test	126.11	−0.1	0.38
Triamcinolone	Train	394.44	1.1	2.48
Triamterene	Train	253.27	1.3	3.39
1,2,3-Trichlorobenzene	Train	181.45	3.8	6.05
Trimethoprim	Train	290.32	0.8	1.71
Uracil	Test	112.09	−1.1	0.49
Uric acid	Train	168.11	−2.4	0.37
Xanthine	Train	152.11	−0.6	0.74

MW, molecular weight; σ, solubilization factor.

* Log *P* calculated using ACD software.

Table II. Statistical Parameters for the Training Set

Model	n	r	rms	F	Q^2
One-group model					
25% PEG	84	0.67	0.83	8.9	0.32
50% PEG	84	0.67	0.79	8.8	0.31
75% PEG	84	0.66	0.75	8.3	0.28
Two-group model					
25% PEG					
Group 1	42	0.92	0.50	25.8	0.62
Group 2	42	0.90	0.49	20.7	0.74
50% PEG					
Group 1	41	0.93	0.39	27.9	0.80
Group 2	43	0.91	0.49	25.1	0.70
75% PEG					
Group 1	43	0.92	0.46	27.3	0.77
Group 2	41	0.89	0.40	18.7	0.66

n, number of compounds; *r,* correlation coefficient; rms, root mean square error (in log units); *F, F*-test value for regression; Q^2, cross-validation correlation coefficient

$$d_{ij} = \Sigma(x_{ik} - x_{jk})^2 \qquad k = 1 - 7 \tag{1}$$

where $x_{i1}, x_{i2}, \ldots x_{ik}$ are the seven molecular descriptors of compound i, $x_{j1}, x_{j2}, \ldots x_{jk}$ are the seven descriptors for compound j, and d_{ij} is the Euclidean distance between compounds i and j. If a smallest Euclidean distance was found between testing set compound A and training set compound B, then solubility of compound A was predicted using the regression model pertaining to compound B's group.

The effectiveness of the model was also evaluated with the cross-validation correlation coefficient (Q^2), the correlation coefficient (r), and the root mean square (rms) error. It was also desired that the model successfully predict solubilities within one order of magnitude (i.e., within 1 log unit). Therefore, the percentages of solubilities predicted within 1.0 log unit (and within 0.5 log unit) were additional indicators used to assess the performance of the model.

Comparison to Log-Linear Model

It was possible to compare the experimental data with the log-linear model for cosolvent solubilization (8). This required the experimental determination of aqueous solubility for all 122 drugs in this study. It should be pointed out that in early drug discovery efforts, such experimental data are not always available due to limited crystalline drug substance.

The log-linear model relates the solubility of a drug in a cosolvent mixture (S_m) to the drug's aqueous solubility (S_w) and a solubilization factor (σ) as follows:

$$\log S_m = \log S_w + \sigma \cdot f \tag{2}$$

where f is the volume fraction of nonaqueous cosolvent in the mixture, and solubility is in mol/l. Furthermore, σ for a drug is related to the drug's log P by the following equation:

$$\sigma = S \cdot \log P + T \tag{3}$$

where S and T are constants specific to a particular cosolvent (10). By obtaining σ values from plots of log S_m vs. f from experimental determinations and then plotting σ vs. log P, one could use the resulting S and T parameters for PEG, a drug's log P, and the solubility of that drug in water to predict solubility in any fraction of PEG.

RESULTS AND DISCUSSION

QSPR Model

The statistical results of the QSPR modeling in Table II demonstrate the benefit of incorporating two groups at each of the three volume fractions of PEG. The regression model based on the solubility values of the training set compounds show r values ranging from 0.89 to 0.93, Q^2 values of 0.62 to 0.80, and rms values of 0.39 to 0.50 log units. The coefficients for the model are listed in Table III, and Fig. 1 shows that the compounds from the training set fit reasonably well to a linear model.

Data from the 84 training set compounds were initially regressed to create the QSPR model. The general form of the QSPR model at each volume fraction of PEG is as follows:

$$\log S = c_0 + c_1 \cdot MW + c_2 \cdot V_m + c_3 \cdot RB + c_4 \cdot HBA + c_5 \cdot HBD + c_6 \cdot RG + c_7 \cdot D_m \tag{4}$$

where S is the solubility (M), MW is the molecular weight

Table III. Coefficients for the QSPR Model*

	Cosolvent					
	25% PEG		50% PEG		75% PEG	
	Group					
Regression coefficients	1	2	1	2	1	2
c_0	1.82E + 01	−5.03E + 00	−3.39E + 00	1.94E + 01	2.07E + 01	−3.96E + 00
c_1	5.01E − 02	−1.26E − 02	8.83E − 03	5.28E − 02	7.22E − 02	−7.35E − 03
c_2	−6.18E − 02	5.80E − 03	−1.86E − 02	−6.32E − 02	−8.26E − 02	−2.38E − 03
c_3	9.75E − 02	2.29E − 01	9.06E − 02	1.92E − 01	1.85E − 01	−6.01E − 03
c_4	−6.38E − 02	−6.27E − 02	−3.11E − 01	−3.02E − 02	−1.11E − 01	−2.92E − 01
c_5	−2.26E − 01	4.04E − 01	4.51E − 01	−2.61E − 01	−2.04E − 01	−2.84E − 01
c_6	−4.65E − 01	−1.61E − 01	4.62E − 01	−7.80E − 01	−3.09E − 01	5.44E − 01
c_7	−1.53E + 01	3.16E + 00	1.20E + 00	−1.55E + 01	−1.77E + 01	3.12E + 00

* Log S = c_0 + c_1·(molecular weight, g/mol) + c_2·(molecular volume, Å^3) + c_3·(number of rotatable bonds) + c_4·(number of hydrogen-bond acceptors) + c_5·(number of hydrogen-bond donors) + c_6·(radius of gyration, Å) + c_7·(molecular density, the ratio of molecular weight/volume).

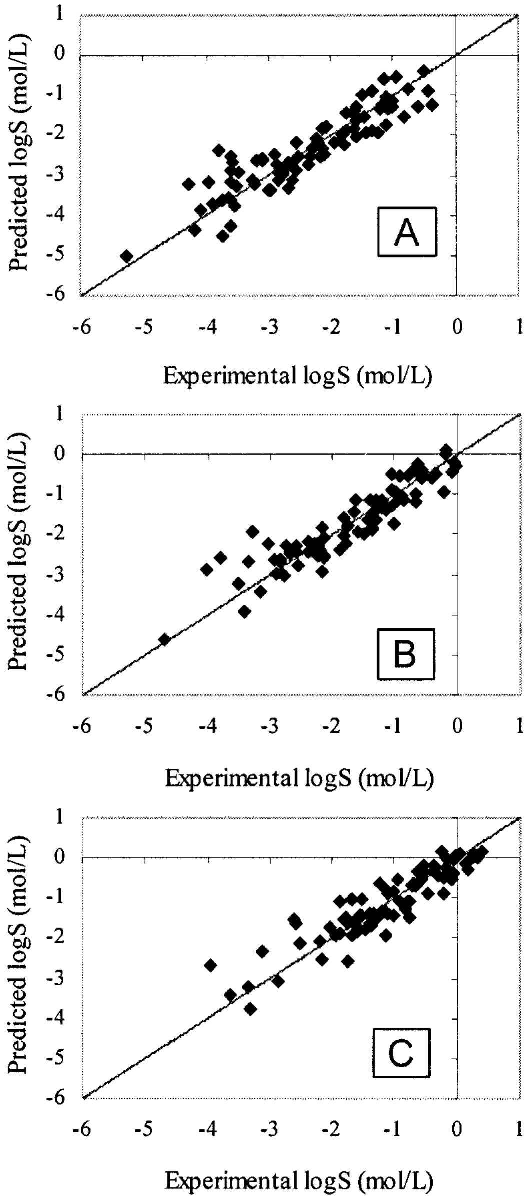

Fig. 1. Training set results. Experimental solubilities vs. predicted solubilities by the QSPR model for (A) 25%, (B) 50%, and (C) 75% PEG.

(g/mol), V_m is the molecular volume (Å^3), RB is the number of rotatable bonds, HBA is the number of hydrogen-bond acceptors, HBD is the number of hydrogen-bond donors, RG is the radius of gyration (Å), and D_m is the molecular density (ratio of molecular weight/volume). The regression coefficients (c_0–c_7) for groups 1 and 2 at each volume fraction of PEG are given in Table III.

The experimental and predicted solubility values of the 38 testing set compounds at each volume fraction were used to assess the predictive ability of the model. As previously described, the 38 testing set compounds were binned into one of two groups based on similarity to the training set, as calculated by Euclidean distance. For the 25% PEG model, 20 compounds were predicted by group 1, and 18 compounds were predicted by group 2. Likewise for the 50% PEG model, 19 compounds were predicted by group 1, and 19 compounds were predicted by group 2. Finally, for the 75% PEG model, 17 compounds were predicted by group 1, and 21 compounds were predicted by group 2. For the testing set compounds, 78.1% and 54.4% of the solubility predictions were within 1.0 and 0.5 log units of observed values, respectively (Table IV).

Table IV shows that a QSPR-predicted testing set solubility value was greater than 2.0 log units from the experimental value in four instances: strychnine and terfenadine in 25% PEG and 5-fluorouracil in both 25% and 50% PEG. Both strychnine and terfenadine were structural outliers in the clustering analysis used to divide the training and testing sets, and they reappear as outliers in the modeling results. While the average Euclidean distance (d_{ij}) for testing set compounds from training set compounds was 0.16, the d_{ij} of terfenadine from its most similar training set compound, clofazimine, was 1.67. 5-fluorouracil was most similar to the training set compound xanthine. Although 5-fluorouracil is about 700 times more soluble in water than xanthine, the computational model does not take aqueous solubility into account, but relies instead upon molecular descriptors. The cosolvent solubilities for 5-fluorouracil were therefore underpredicted to be similar to xanthine.

The training and testing sets were divided using both cluster analysis and random selection. This selection method was repeated to generate a second model with different training and testing sets containing 82 and 40 compounds, respectively. This model predicted 77.5% of its testing set solubilities within 1.0 log unit (compared to 78.1% for the original

Table IV. Residual Distributions for Predicted Testing Set Solubilities

	Residual ranges in log units				
	< ± 0.5	± 0.5 to ± 1.0	± 1.0 to ± 1.5	± 1.5 to ± 2.0	> ± 2.0
QSPR model					
25% PEG	18* (47.4%)	12 (31.6%)	3 (7.9%)	2 (5.3%)	3 (7.9%)
50% PEG	24 (63.2%)	6 (15.8%)	5 (13.2%)	2 (5.3%)	1 (2.6%)
75% PEG	20 (52.6%)	9 (23.7%)	4 (10.5%)	5 (13.2%)	0 (0%)
Total	62 (54.4%)	27 (23.7%)	12 (10.5%)	9 (7.9%)	4 (3.5%)
Log-linear model					
25% PEG	29 (76.3%)	6 (15.8%)	2 (5.3%)	1 (2.6%)	0 (0%)
50% PEG	22 (57.9%)	13 (34.2%)	2 (5.3%)	1 (2.6%)	0 (0%)
75% PEG	14 (36.8%)	12 (31.6%)	8 (21.1%)	2 (5.3%)	2 (5.3%)
Total	65 (57%)	31 (27.2%)	12 (10.5%)	4 (3.5%)	2 (1.8%)

* Number of compounds (percentage) having the predicted solubility within a specified number of log units of the experimental value.

model), serving as proof of the model's robustness. It is interesting to note that this second model also resulted in four outlying residuals greater than 2.0 log units.

In recent years, *in silico* quantitative structure–property relationships have been used for the prediction of various physicochemical properties. A QSPR is a mathematical relationship between a property of interest and structural characteristics of the compounds. The structural features of a compound are quantified by a series of molecular descriptors that encode the topological, geometric, and electronic information of the molecule. Selection of appropriate molecular descriptors and use of an appropriate data set is critical to model success (16). Significant advances have been made in the development of new molecular descriptors based on recent advances in computational abilities. In many of these studies, complex molecular descriptors such as E-state indices, total weighted number of paths, and cube root of gravitation index were used, making the interpretation of the model difficult (17,18). In the current analysis, descriptor selection was limited to those that are physically and chemically relevant and can directly be related to solubility of drugs in PEG. By applying some of the more advanced predictors, it may be possible to improve model predictiveness. While many QSPR models have been developed for predicting aqueous solubility (13), there are no reports in the literature which apply QSPR models for predicting drug solubility in cosolvent systems.

Log-Linear Model

For comparison of the QSPR modeling results to the log-linear model, σ was calculated from the experimental data in 0%, 25%, 50%, and 75% PEG as described by Millard et al. (12). *S* and *T* parameters obtained from the training set σ values were used with experimental aqueous solubility data and log *P* values for the testing set compounds to predict cosolvent solubility values for the testing set.

Table IV and Fig. 2 show that the QSPR and log-linear models predict the testing set solubility values with comparable accuracy. Considering the theoretical basis of the log-linear model, it is not surprising that this model predicts solubilities in 25% PEG better than in the higher volume fractions and better than the QSPR model in 25% PEG. One would expect the solubility at a low fraction of cosolvent to be close to its aqueous solubility, an experimental input for the log-linear prediction.

Although the predictive abilities of the QSPR and log-linear models are similar, a key difference between the two methods is that the QSPR model requires no experimental data—the cosolvent solubility values are calculated from molecular descriptors obtained from a chemical structure. Prediction by the log-linear equation, on the other hand, requires the experimental determination of aqueous solubility.

Several authors have observed positive deviation from log-linear behavior at higher fractions of cosolvent (3,19–23). In this study, solubility was measured across the entire PEG fraction range for 94 drugs, and about half of these compounds displayed such deviation. Rubino et al. (21,22) suggest that hydrogen-bond donating groups can compete with water for hydrogen-bond accepting sites on the cosolvent molecule. Consistent with the suggestion, we observed that the magnitude of a compound's deviation from log-linearity at high fractions of PEG correlated well with the number of hydrogen-bond donors for compounds containing no phenyl rings. However, for the drugs containing phenyl rings, the lack of a strong correlation may be due to steric interference of the phenyl rings with hydrogen bonding (14) or the existence of oxygen-aromatic interactions with the cosolvent (24).

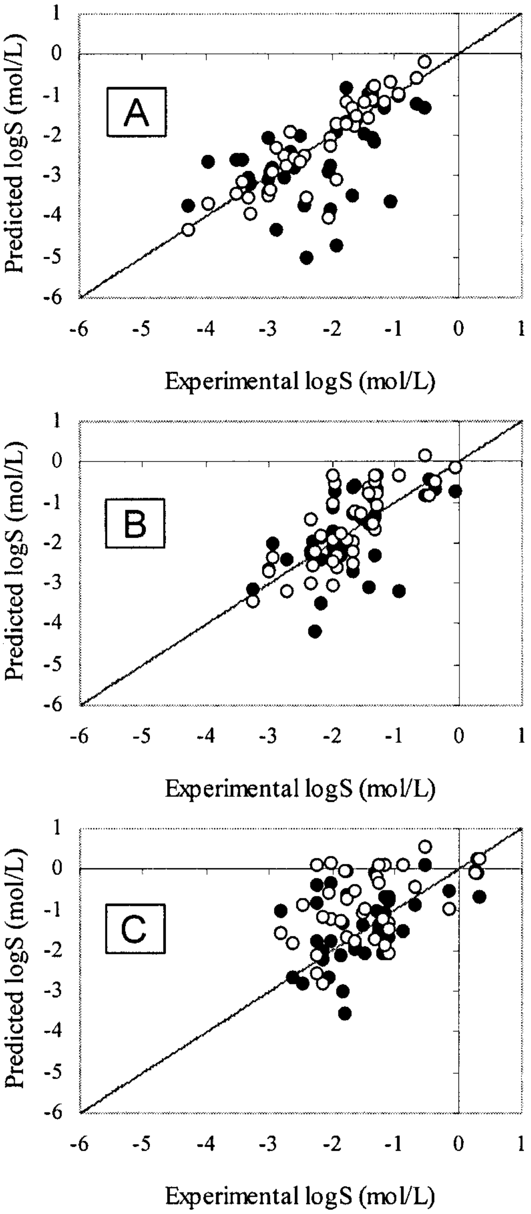

Fig. 2. Testing set results. Experimental solubilities vs. predicted solubilities by the QSPR model (closed circles) and the log-linear model (open circles) for (A) 25%, (B) 50%, and (C) 75% PEG.

Despite these deviations from ideality, the log-linear model was still a useful predictive tool. The regression of *S* and *T* parameters from 122 experimental σ values in this study appears in Fig. 3. Even though there is fair agreement between σ values for individual drugs overlapping this work and Ref. 12, the *S*, *T*, and *r* values in Table V are not in complete accord, perhaps due to the larger compound base used in this study. Whereas the regression of PEG 400 log-linear *S* and *T* parameters in this study gave $r = 0.67$ (n = 122), Millard et al. report the *r* for σ and log *P* in ethanol to

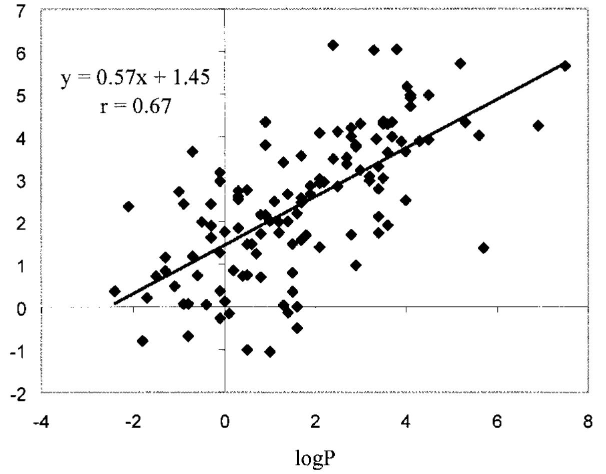

Fig. 3. Regression of log-linear model *S* and *T* parameters for the experimental data collected in this work (n = 122).

be 0.98 (n = 120) (12). This suggests that solubilization power in PEG does not correlate with log *P* as well as solubilization in a more octanol-like cosolvent, such as ethanol; this observation needs further investigation.

CONCLUSIONS

A QSPR model for predicting solubility of drugs in three volume fractions of PEG was developed based on experimental data. This tool for vehicle identification during lead optimization can be used to predict solubility of a compound in a PEG/water cosolvent mixture from the knowledge of its chemical structure alone, without experimental measurements. This model is especially useful during early drug discovery because no compound is required, thus making the prediction fast and practical. Often, the pharmaceutical scientist has minimal amounts of compound (5–10 mg), and it is desired to identify a preclinical formulation at a target concentration of 10–20 mg/ml; hence any method(s) that can guide vehicle selection would be valuable. In the interest of keeping the model simple and with the long-term objective of extending *in silico* predictions into new areas, the model was limited to seven descriptors. This model will be applied next to the prediction of cosolvent solubility of internal Bristol-Myers Squibb compounds in various PEG/water fractions. An immediate advantage has been the rapid identification of appropriate cosolvent systems necessary for early preclinical animal studies.

When aqueous solubility of crystalline drug substance is available, log-linear estimation is simple and effective. The log-linear model predicted the testing set solubilities as well as the QSPR model, but the log-linear model is dependent on an experimental measurement of aqueous solubility. Thus, the QSPR model described here benefits pharmaceutical scientists at the discovery phase seeking a quick and reliable estimation of cosolvent solubility, without expending compound.

Table V. Comparison of Log-Linear Model *S* and *T* Parameters for PEG 400

Source	n	Log *P* range	*S*	*T*	*r*
Ref. 12	25	−2.4 to 6.8	0.74 ± 0.07	1.26 ± 0.22	0.91
This work	122	−2.4 to 7.5	0.57 ± 0.06	1.45 ± 0.15	0.67

n, number of compounds; r, correlation coefficient.

ACKNOWLEDGMENTS

The authors would gratefully like to acknowledge Drs. Greg Pearl and Stephen Johnson for their insightful review and comments and Drs. Atsu Apedo and Ning Zhao for analytical assistance. E. R. was a summer intern at Bristol-Myers Squibb Pharmaceutical Research Institute.

REFERENCES

1. S. Venkatesh and R. A. Lipper. Role of the development scientist in compound lead selection and optimization. *J. Pharm. Sci.* **89**: 145–154 (2000).
2. S. Sweetana and M. J. Akers. Solubility principles and practices for parenteral drug dosage form development. *PDA J. Pharm. Sci. Technol.* **50**:330–342 (1996).
3. S. H. Yalkowsky. Solubilization by cosolvents. In S. H. Yalkowsky (ed.), *Solubility and Solubilization in Aqueous Media*, Oxford University Press, New York, 1999 pp. 180–235.
4. Final Report on the Safety Assessment of Polyethylene Glycols. (PEGs) -6, -8, -32, -75, -150, -14M, -20M. *J. Am. Coll. Toxicol.* **12**:429–457 (1993).
5. A. Jouyban-Gharamaleki, L. Valaee, M. Barzegar-Jalali, B. J. Clark, and W. E. Acree, Jr. Comparison of various cosolvency models for calculating solute solubility in water-cosolvent mixtures. *Int. J. Pharm.* **177**:93–101 (1999).
6. N. A. Williams and G. L. Amidon. Excess free energy approach to the estimation of solubility in mixed solvent systems I: theory. *J. Pharm. Sci.* **73**:9–13 (1984).
7. A. Martin, A. N. Paruta, and A. Adjei. Extended Hildebrand solubility approach: methylxanthines in mixed solvents. *J. Pharm. Sci.* **70**:1115–1120 (1981).
8. S. H. Yalkowsky, G. L. Flynn, and G. L. Amidon. Solubility of nonelectrolytes in polar solvents. *J. Pharm. Sci.* **61**:983–984 (1972).
9. A. Li and S. H. Yalkowsky. Predicting cosolvency. 1. Solubility ratio and solute logKow. *Ind. Eng. Chem. Res.* **37**:4470–4475 (1998).
10. S. H. Yalkowsky and T. J. Roseman. Solubilization of drugs by cosolvents. In S. H. Yalkowsky (ed.), *Techniques of Solubilization of Drugs.*, Marcel Dekker, New York, 1981, pp. 91–134.
11. S. H. Yalkowsky, S. C. Valvani, and G. L. Amidon. Solubility of nonelectrolytes in polar solvents IV: nonpolar drugs in mixed solvents. *J. Pharm. Sci.* **65**:1488–1494 (1976).
12. J. W. Millard, F. A. Alvarez-Núñez, and S. H. Yalkowsky. Solubilization by cosolvents: Establishing useful constants for the log-linear model. *Int. J. Pharm.* **245**:153–166 (2002).
13. X-Q. Chen, S. J. Cho, Y. Li, and S. Venkatesh. Prediction of aqueous solubility of organic compounds using a quantitative structure-property relationship. *J. Pharm. Sci.* **91**:1838–1852 (2002).
14. D. J. W. Grant and T. Higuchi. Solubility, intermolecular forces, and thermodynamics. In *Solubility Behavior of Organic Compounds*, John Wiley & Sons, New York, 1990, pp. 12–88.
15. S. J. Cho and M. A. Hermsmeier. Genetic algorithm guided selection: variable selection and subset selection. *J. Chem. Inf. Comput. Sci.* **42**:927–936 (2002).
16. T. R. Stouch, J. R. Kenyon, S. R. Johnson, X. Q. Chen, A. Doweyko, and Y. Li. In silico ADME/Tox: why models fail. *J. Comput. Aided Mol. Des.* **17**:83–92 (2003).
17. I. V. Tetko. V. Y Tanchuk, T. N. Kasheva, and A. E. P. Villa. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **41**:1488–1493 (2001).
18. B. E. Mitchell and P. C. Jurs. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **38**:489–496 (1998).
19. E. Khalil, S. Najjar, and A. Sallam. Aqueous solubility of diclofenac diethylamine in the presence of pharmaceutical additives: a

comparative study with diclofenac sodium. *Drug Dev. Ind. Pharm.* **26**:375–381 (2000).

20. T. A. Hagen and G. L. Flynn. Solubility of hydrocortisone in organic and aqueous media: evidence for regular solution behavior in apolar solvents. *J. Pharm. Sci.* **72**:409–414 (1983).
21. J. T. Rubino and S. H. Yalkowsky. Cosolvency and deviations from log-linear solubilization. *Pharm. Res.* **4**:231–236 (1987).
22. J. T. Rubino and E. K. Obeng. Influence of solute structure on deviations from the log-linear solubility equation in propylene glycol:water mixtures. *J. Pharm. Sci.* **80**:479–483 (1991).
23. R. Tarantino, E. Bishop, F. C. Chen, K. Iqbal, and A. W. Malick. N-methyl-2-pyrrolidone as a cosolvent: relationship of cosolvent effect with solute polarity and the presence of proton-donating groups on model drug compounds. *J. Pharm. Sci.* **83**:1213–1216 (1994).
24. S. K. Burley and G. A. Petsko. Weakly polar interactions in proteins. *Adv. Protein Chem.* **39**:125–189 (1988).