Development of Neural Network QSPR Models for Hansch Substituent Constants. 1. Method and Validations

Ting-Lan Chiu and Sung-Sau So*

Roche Research Center, Hoffmann-La Roche Inc., Nutley, New Jersey 07110

Received June 24, 2003

In an attempt to develop predictive models for Hansch substituent constants for less common substituents, neural network QSPR (Quantitative Structure−Property Relationship) studies were conducted to correlate Hansch substituent constants for hundreds of chemically diverse functional groups with two different molecular descriptor sets. The Hansch substituent constants under study were π, MR, F and R, describing the hydrophobic, steric/polarizability, and electronic (field and resonance) characteristics of the substituents, respectively. E-state descriptors were used for π and MR, while the molecular descriptor set based upon the approach of Kvasnicka, Sklenak, and Pospichal (*J. Am. Chem. Soc.* **1993**, *115*, 1495−1500) was adopted for F and R. Both QSPR models demonstrated good predictivity in test sets.

INTRODUCTION

QSAR (Quantitative Structure−Activity Relationship) correlates biological activity data with the physiochemical and/or structural properties of a group of compounds. It has been frequently used to predict biological activities of new compounds and to design compounds with desired properties. The most relevant properties of a compound to drug design are hydrophobic, steric, and electronic properties,[1] since they are the primary forces of molecular recognition. In particular, π (the Hansch-Fujita π constant) is frequently used as a measure of the hydrophobic effect of a substituent.[2−6] The commonly used constant to capture steric and polarizability effect of a substituent is MR (molar refractivity). F (Swain and Lupton field parameter) and R (Swain and Lupton resonance parameter) are field/inductive and resonance components, respectively, of the Hammett σ constants that are generally used to quantify the electronic effects of substituents. Detailed definitions of these substituent constants have been published.[7]

Despite the fact that standard values of Hansch substituent constants have been tabulated for hundreds of substituents,[8,9] there are still many common substituents for which parameters are not available. For example, standard π, MR, F and R values of $-NHCH_3$, $-NHC_2H_5$ and $-N(CH_3)_2$ can be found, but those of $-N(C_2H_5)_2$ cannot.[8,9] As a result, compounds with functional groups lacking standard substituent constants must be removed from any analysis that utilizes such constants. Despite this limitation, few studies have been undertaken to predict Hansch substituent constants. To our knowledge, only two studies have specifically addressed this issue.[10,11] Instead, various approaches have been proposed previously to predict whole molecule properties that are highly related to substituent constants. Since the approaches used to model these molecular properties could very likely be applied to model substituent constants, it is useful to review these previous studies here. As an example, the Hansch-Fujita π substituent constant is related to another commonly accepted hydrophobic measure, log *P* values[12] which have been derived by two approaches: atom types[13] and hydrophobic fragmental constants.[14] Recently, Tetko and co-workers have also used E-state descriptors combined with artificial neural networks to predict log *P* values.[15]

MR has also been predicted with high accuracy by the atom-type approach. For example, Ghose and Crippen applied a constrained least-squares technique on 504 compounds to derive atomic refractivities of 93 atom-types, with a standard deviation of 1.269 and a correlation coefficient of 0.994.[16] The atomic refractivities were then combined to predict molar refractivities of 78 compounds. The standard deviation and correlation coefficient for prediction were 1.164 and 0.994, respectively. Viswanadhan and co-workers extended this concept and employed a technique called quadratic programming that allows for least-squares fitting of molar refractivities while constraining individual atomic contributions to reside within a physically relevant region. They determined the atomic refractivities of 120 atom-types, with a standard deviation of 0.774 and a correlation coefficient of 0.998[17] and then used these atomic refractivities to predict molar refractivities of 82 compounds, yielding a standard deviation of 1.553 and a correlation coefficient of 0.996.

A summary of previous approaches for determining electronic substituent constants (e.g., F, R, and σ) is shown in Table 1.[10,11,18−21] Some are regression-based models that fit substituent constants to electronic parameters obtained from quantum mechnical/semiempirical calculations or electrostatic properties on a grid lattice. Obviously, this type of calculation requires the use of three-dimensional structures. There are also methods that depend solely on two-dimensional topological descriptors. All in all, most of these studies tackled data sets consisting only of common, simple substituents without charges, complicated cyclic substructures or heterocycles (except pyridine,[21] thiophene,[21] and naphthalene[18]).

* Corresponding author phone: (973)235-2193; fax: (973)235-2682; e-mail: sung-sau.so@roche.com.

10.1021/ci030293q CCC: $27.50
Published on Web 12/16/2003

Table 1. A Comparison of Approaches to Deriving Various Electronic Parameters: F, R and σ

substituent descriptors	algorithm	substituent constant(s)	ref
graph-theoretical parameters	neural networks	F, R	11
electrostatic field values	CoMFA/PLS	F, R	19
molecular connectivity δ values	linear regression	σ	20
frontier orbital energies	linear regression	σ	10
semiempirical atomic charges	linear regression	σ	21
quantum self-similarity measures	linear regression	σ	18

The goal of this study is to develop Quantitative Structure−Property Relationship (QSPR) models that accurately predict the four substituent constants, π, MR, F and R. Ideally, the models would handle diverse structural variations (e.g. charged groups or heterocyclic aromatic systems), while the calculations should be fast and easy to automate, thereby overcoming the major limitations of previous approaches. These goals were achieved by the following strategy: (1) using topological (i.e. two-dimensional and conformation-independent) descriptors for all four substituent constants; (2) applying an artificial neural network algorithm to correlate substituent constants with descriptors; and (3) including substituents of diverse structural classes in both training and test sets. Specifically, neural network QSPR studies were conducted to correlate π, MR, F and R with two different molecular descriptor sets for hundreds of chemically diverse substituents. For π and MR, E-state descriptors were used for correlation. The E-state descriptors were selected because they have been used in establishing numerous QSPR and QSAR models.[15,22−25] For F and R, Kvasnicka, Sklenak and Pospichal[11] (abbreviated hereafter as KSP) developed a set of graph-theoretical descriptors and reported very impressive prediction models when the descriptors were used in conjugation with artifical neural networks. These KSP descriptors have been further extended in this work and used for the same purpose. Both the E-state and KSP descriptor sets were two-dimensional in nature and were thus simple to implement, fast to calculate, and independent of molecular conformation. The QSPR models demonstrated good predictivity in all our test sets and leave-one-out cross-validation procedures.

DETAILS OF MODEL DEVELOPMENT

Data Sets. The π, MR, F, and R values of hundreds of substituents were taken from refs 8 and 9. The following substituents were removed from our data set: (1) substituents containing the following metal elements: Sn, As, Cr, Co, Mn, Fe, Te, Ge, Mo, Hg and (2) substituents forming a fused ring with the main group, e.g, —OCH_2O—, leaving us with a data set containing 764 substituents. This complete data set is available in the Supporting Information for interested readers. It is important to note that all four constants were not available for these 764 substituents. Specifically, there are 327 π values, 602 MR values, and 490 F and R values. For each substituent constant, the corresponding data set was randomly divided into training and test sets in 4:1 ratio. The number of substituents in the training set versus the test set is 262 versus 65 for π, 482 versus 120 for MR, and 392 versus 98 for both F and R.

Descriptor Selection. The descriptor set for π contains a total of 19 descriptors: 18 composite E-state descriptors[26] and the charge of the substituent. The charge is considered to be an important descriptor for π since the higher the charge, the lower the hydrophobicity. The value of each individual E-state descriptor was first determined by summing up the E-state values of the atoms that belong to the corresponding atom-type.[26] Similar types of E-state descriptors were then combined to ensure that each of the final 18 composite E-state descriptors (ES_{0-17}) was sufficiently represented among the substituents in our data set. This avoided sparse or skewed distributions. For example, all five E-state atom-types of iodine occurring in the data set are combined to a single composite descriptor, ES_{11}. Table 2 provides the details for the 18 composite E-state descriptors. For MR, the descriptor set consists of the same 18 composite E-state descriptors as well as the molecular weight of the substituent. The molecular weight is included due to its correlation with the size of the substituent.

Table 2. The E-State Descriptors Used To Model π and MR[a]

ES_0	SssssC + SsssCH + SssCH2 + SsCH3
ES_1	(StC-) + StsC + SddC + SdssC + (SsssC+) + StCH + SdsCH + SdCH2
ES_2	SaasC + SaaCH
ES_3	SsssN + (SssssN+) + SssNH + SsNH2 + (SssNH2+) + (SsNH3+) + (SsssNH+)
ES_4	StN + (SdN-) + SdsN + (StsN+) + (SddN+) + SddsN + SdNH
ES_5	SaaN + SaasN + SaadN + SaaNH
ES_6	(SsO-)+SssO+SsOH
ES_7	SdO + SaaO
ES_8	SsF
ES_9	SsCl
ES_{10}	SsBr
ES_{11}	SsI + SsssI + SddsI + SsssssI + SdsI
ES_{12}	SsSiH3 + SssSiH2 + SsssSiH + SssssSi
ES_{13}	(SssssP+) + SssssssP + SsPH2 + SssPH + SsssP
ES_{14}	SdsssP
ES_{15}	SsSH + SssS + SsssssS + SssssssS + (SsssS+) + (SsS-)
ES_{16}	SdS + SaaS + SdssS + SddssS + SdaaS
ES_{17}	SssSe + SddssSe + SdssSe + SsSeH + SdSe + SaaSe

[a] These E-state atom-types are combined into 18 (0−17) descriptors, ES_0−ES_{17}.

The descriptor sets for F and R were modified graph-theoretical parameters derived originally by Kvasnicka, Sklenak and Pospichal.[11] Briefly, the KSP descriptors take the first-, second- and third-shell atoms of the substituent into consideration. The first-shell atom is the atom at the attachment point. The second-shell atoms are directly attached to the first-shell atom and the third-shell atoms to the second-shell atoms. The following properties are calculated for the first-shell atoms: (1) the number of electron lone pairs, (2) the quantum number minus 1, (3) the number of hydrogen atoms attached to this atom, and (4) the number of phenyl groups attached to this atom. For the atoms in the second and the third shells, the number of π bonds is added as the fifth descriptor. The KSP descriptors were modified for this study as follows: (1) The descriptors for the fourth-shell atoms were added in order to account for longer-range contributions to the electronic effect and (2) the total formal charge for each shell was also added to the descriptor set. The second modification was considered necessary, since the lack of consideration for charged and heterocyclic groups is a limitation of the KSP approach. In addition, instead of counting the number of phenyl groups at each shell, the number of aromatic atoms was counted. To sum up, there

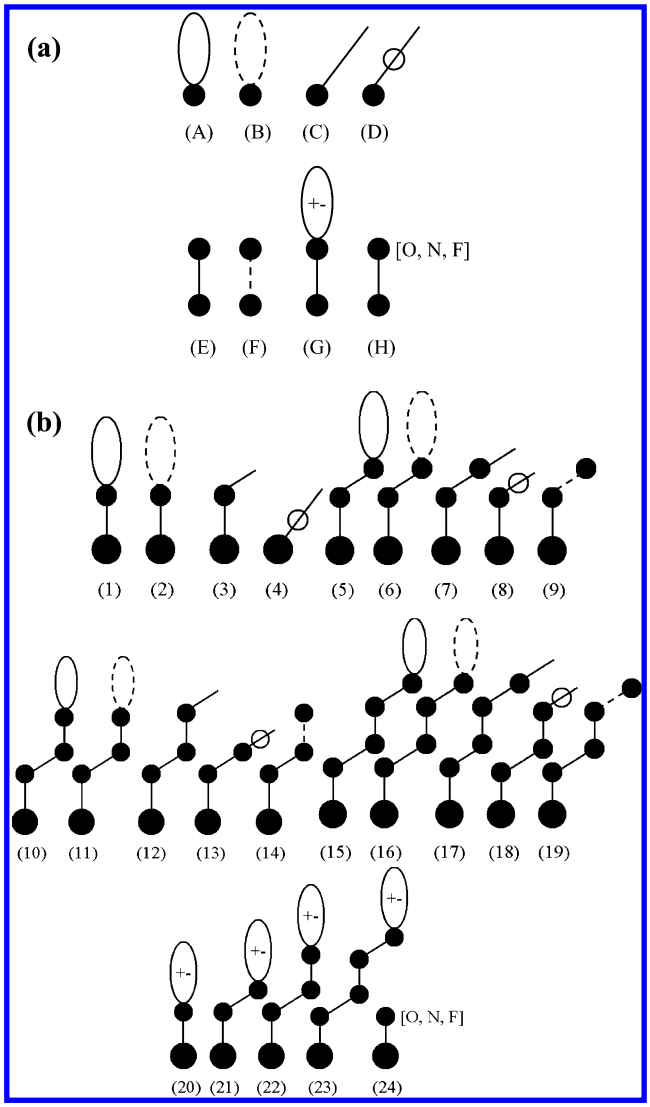

Figure 1. (a). Symbols used to construct the molecular subgraphs in Figure 1(b): (A) the lone electron pair, (B) the main quantum number minus 1, (C) the hydrogen atom attached to an atom, (D) the number of aromatic atoms in the aromatic ring attached to an atom, (E) the σ bond, (F) the π bond, (G) the total formal charge for a shell, and (H) indicator descriptor (equal to 1 if the atom is O, N, or F and equal to 0, otherwise). (b). The molecular subgraphs for the descriptors used for F and R. F has 23 descriptors illustrated in (1)−(23) while R used 24 descriptors illustrated in (1)−(24). The bold dots (roots) are the atoms that are immediately attached to an atom of the substituent.

were a total of 23 descriptors (14 modified KSP descriptors + 5 fourth-shell descriptors + 4 total formal charges for four shells) for F. For R, one more indicator descriptor was added to emphasize the resonance effect from the first-shell atom. The descriptor has a value 1 if the first-shell atom is O, N or F (with lone pairs) and 0 otherwise. Consequently, the descriptor sets for F and R numbered 23 and 24, respectively. The molecular subgraphs for these 24 descriptors were constructed following Kvasnicka, Sklenak, and Pospichal,[11] as can be seen in Figure 1. Table 3 details the descriptor values of F and R for substituents, NO_2, CH_2Ph, SO_2Ph, and $NHCOCH_3$, which have also been used as examples by Kvasnicka, Sklenak, and Pospichal.[11] All descriptors were generated by an in-house program that utilizes Daylight toolkit libraries.[27]

Table 3. Descriptor Values of F and R for Four Example Molecules[a]

X	NO_2	CH_2Ph	SO_2Ph	$NHCOCH_3$
x_1	0	0	0	1
x_2	1	1	2	1
x_3	0	2	0	1
x_4	0	0	0	0
x_5	4	0	4	0
x_6	2	1	3	1
x_7	0	0	0	0
x_8	0	1	1	0
x_9	2	0	2	0
x_{10}	0	0	0	2
x_{11}	0	2	2	2
x_{12}	0	2	2	3
x_{13}	0	2	2	0
x_{14}	0	0	0	1
x_{15}	0	0	0	0
x_{16}	0	2	2	0
x_{17}	0	2	2	0
x_{18}	0	2	2	0
x_{19}	0	0	0	0
x_{20}	0	0	0	0
x_{21}	0	0	0	0
x_{22}	0	0	0	0
x_{23}	0	0	0	0
x_{24}	0	0	0	1

[a] Figure 1 gives a graphical illustration of descriptors x_1−x_{24}.

Regression Method. In an effort to explore nonlinear dependencies of Hansch substituent constants on the descriptors, a three-layered, fully connected neural network was constructed on the training set. To find the optimal number of neurons in the hidden layer, a neural network training with 19:h:1, 19:h:1, 23:h:1, and 24:h:1 architectures, where $h = 2$ to 6, was carried out for π, MR, F and R, respectively. The optimal number of hidden neurons was found to be 3 for all four cases on the basis of the correlation coefficients and standard deviations between the predicted and standard tabulated values of the training set.

RESULTS AND DISCUSSION

The goal of this study was to generate efficient QSPR models that can reliably predict substituent constants for less common type of fragments. The use of substituent constants to characterize the biological activity of congeneric series has been a cornerstone of many structure−activity relationships for over forty years. In our opinion, there are two major advantages of using substituent constants (as compared to topological descriptors such as E-state or KSP) for QSAR modeling. First, the substituent constants are physically meaningful, and therefore the resulting QSAR models are often easier to interpret. Second, because only a relatively few descriptors (e.g. π, MR, F and R) are needed to characterize each variable substituent position, the potential risk of model overfitting is greatly reduced.

Table 4 lists the important statistical figures for the training set, test set and the leave-one-out cross-validation approach. The relevant plots showing the correlation between predicted versus standard Hansch substituent constants of the training set and test set are illustrated in Figures 2−5. It is evident that the QSPR models are generally robust, and they all yield very satisfactory prediction results. Of the four substituent constants, the best correlation coefficients were achieved for MR with the full descriptor set: 0.98, 0.92 and 0.96 for the

Table 4. Statistical Parameters for Training Set, Test Set and Leave-One-Out Cross-Validation[a]

π	training	test	CV
r^2/q^2	0.84	0.84	0.81
rms	0.58	0.57	0.63
N	262	65	262
max	3.61	3.26	3.61
min	−4.82	−3.85	−4.82
MR	**training**	**test**	**CV**
r^2/q^2	0.98	0.92	0.96
rms	0.18	0.30	0.24
N	482	120	482
max	6.74	7.02	6.74
min	0.09	0.28	0.09
F	**training**	**test**	**CV**
r^2/q^2	0.88	0.72	0.77
rms	0.08	0.13	0.11
N	392	98	392
max	1.18	1.16	1.18
min	−0.11	−0.07	−0.11
R	**training**	**test**	**CV**
r^2/q^2	0.84	0.71	0.64
rms	0.08	0.11	0.12
N	392	98	392
max	0.47	0.42	0.47
min	−0.77	−0.79	−0.77

[a] r^2: Pearson correlation coefficient, q^2: leave-one-out cross-validated correlation coefficient. rms: root-mean-square deviation. N: number of compounds in the data set. max: maximal value of all substituents in the data set. min: minimal value of all substituents in the data set.

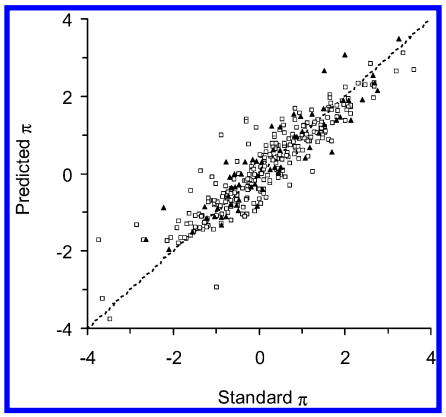

Figure 2. Correlation between predicted versus standard π values of the training set substituents (open squares) and test set substituents (filled triangles). The dotted line represents a perfect correlation between predicted and standard π.

training, test and leave-one-out, respectively. The full descriptor sets for F and R predicted the test set data with reasonable accuracy, given the fact that they were two-dimensional in nature. An outlier analysis was conducted to identify the outliers in the leave-one-out cross-validated sets of F and R. For F, one outlier $-SO_2^-$, which contributed 6.7% of the mean square error (MSE), was identified. Removing this outlier led to an increase of q^2 from 0.765 to 0.780. For R, two outliers, $-NHCH_2SO_3^-$ and $-OCH_2-CH_2O^-$, contributed 16.3% and 5.9% of the MSE of the leave-one-out cross-validated set, respectively. The removal of the first outlier ($-NHCH_2SO_3^-$) yielded a significant increase of q^2 from 0.642 to 0.693; and with both outliers removed, q^2 became 0.715. The results of this analysis

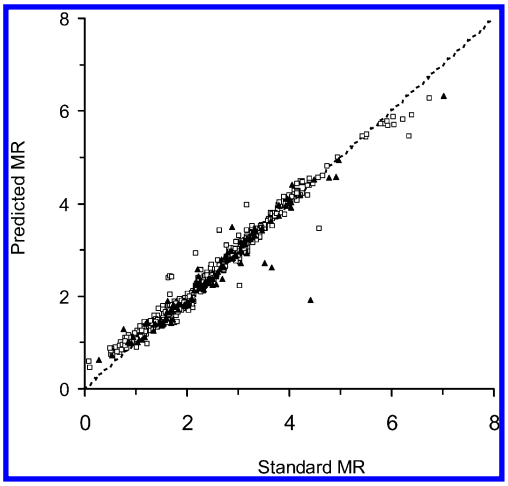

Figure 3. Correlation between predicted versus standard MR values of the training set substituents (open squares) and test set substituents (filled triangles). The dotted line represents a perfect correlation between predicted and standard MR.

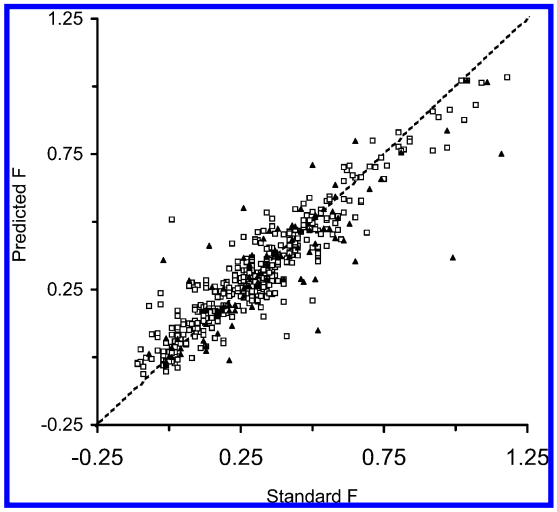

Figure 4. Correlation between predicted versus standard F values of the training set substituents (open squares) and test set substituents (filled triangles). The dotted line represents a perfect correlation between predicted and standard F.

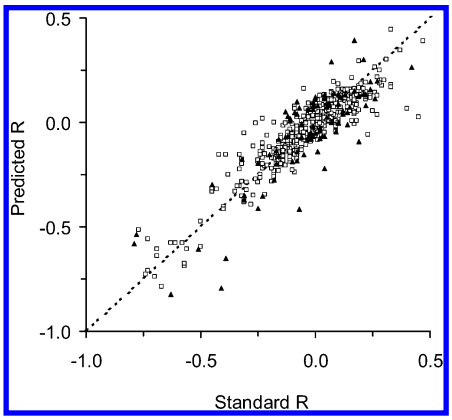

Figure 5. Correlation between predicted versus standard R of the training set substituents (open squares) and test set substituents (filled triangles). The dotted line represents a perfect correlation between predicted and standard R.

indicated that the QSPR models for F and R are generally quite accurate for neutral and positively charged substituents.For negatively charged functional groups, one may encounter larger prediction errors.

Table 5. Results of Sensitivity Analysis[a]

Trn/Tst	training	test	CV	full	training	CV	Trn/Tst	training	test	CV	full	training	CV
			π (1)							π (2)			
r2/q2	0.839	0.837	0.810	r2/q2	0.843	0.815	r2/q2	0.678	0.733	0.597	r2/q2	0.717	0.637
rms	0.582	0.566	0.631	rms	0.569	0.618	rms	0.822	0.724	0.918	rms	0.764	0.865
N	262	65	262	N	327	327	N	262	65	262	N	327	327
max	3.61	3.26	3.61	max	3.61	3.61	max	3.61	3.26	3.61	max	3.61	3.61
min	−4.82	−3.85	−4.82	min	−4.82	−4.82	min	−4.82	−3.85	−4.82	min	−4.82	−4.82
			MR (1)							MR (2)			
r2/q2	0.983	0.920	0.970	r2/q2	0.970	0.957	r2/q2	0.942	0.832	0.918	r2/q2	0.927	0.895
rms	0.153	0.305	0.200	rms	0.198	0.238	rms	0.280	0.441	0.332	rms	0.309	0.371
N	482	120	482	N	602	602	N	482	120	482	N	602	602
max	6.74	7.02	6.74	max	7.02	7.02	max	6.74	7.02	6.74	max	7.02	7.02
min	0.09	0.28	0.09	min	0.09	0.09	min	0.09	0.28	0.09	min	0.09	0.09
			F (1)							F (2)			
r2/q2	0.875	0.724	0.765	r2/q2	0.819	0.75	r2/q2	0.520	0.491	0.469	r2/q2	0.633	0.498
rms	0.082	0.133	0.112	rms	0.101	0.119	rms	0.161	0.181	0.169	rms	0.144	0.168
N	392	98	392	N	490	490	N	392	98	392	N	490	490
max	1.18	1.16	1.18	max	1.18	1.18	max	1.18	1.16	1.18	max	1.18	1.18
min	−0.11	−0.07	−0.11	min	−0.11	−0.11	min	−0.11	−0.07	−0.11	min	−0.11	−0.11
			F (3)							F (4)			
r2/q2	0.504	0.513	0.435	r2/q2	0.589	0.481	r2/q2	0.834	0.682	0.739	r2/q2	0.808	0.722
rms	0.164	0.177	0.174	rms	0.152	0.171	rms	0.095	0.143	0.106	rms	0.090	0.111
N	392	98	392	N	490	490	N	392	98	392	N	490	490
max	1.18	1.16	1.18	max	1.18	1.18	max	1.18	1.16	1.18	max	1.18	1.18
min	−0.11	−0.07	−0.11	min	−0.11	−0.11	min	−0.11	−0.07	−0.11	min	−0.11	−0.11
			R (1)							R (2)			
r2/q2	0.840	0.706	0.642	r2/q2	0.785	0.737	r2/q2	0.710	0.615	0.554	r2/q2	0.671	0.573
rms	0.083	0.114	0.122	rms	0.110	0.122	rms	0.111	0.131	0.137	rms	0.119	0.135
N	392	98	392	N	490	490	N	392	98	392	N	490	490
max	0.47	0.42	0.47	max	1.18	1.18	max	0.47	0.42	0.47	max	0.47	0.47
min	−0.77	−0.79	−0.77	min	−0.11	−0.11	min	−0.77	−0.79	−0.77	min	−0.79	−0.79
			R (3)							R (4)			
r2/q2	0.693	0.490	0.592	r2/q2	0.753	0.625	r2/q2	0.738	0.602	0.630	r2/q2	0.725	0.622
rms	0.114	0.150	0.132	rms	0.103	0.127	rms	0.106	0.133	0.125	rms	0.108	0.127
N	392	98	392	N	490	490	N	392	98	392	N	490	490
max	0.47	0.42	0.47	max	0.47	0.47	max	0.47	0.42	0.47	max	0.47	0.47
min	−0.77	−0.79	−0.77	min	−0.79	−0.79	min	−0.77	−0.79	−0.77	min	−0.79	−0.79
			R (5)							R (6)			
r2/q2	0.700	0.571	0.620	r2/q2	0.754	0.665	r2/q2	0.766	0.706	0.625	r2/q2	0.753	0.646
rms	0.113	0.138	0.101	rms	0.103	0.120	rms	0.100	0.114	0.107	rms	0.103	0.123
N	392	98	392	N	490	490	N	392	98	392	N	490	490
max	0.47	0.42	0.47	max	0.47	0.47	max	0.47	0.42	0.47	max	0.47	0.47
min	−0.77	−0.79	−0.77	min	−0.79	−0.79	min	−0.77	−0.79	−0.77	min	−0.79	−0.79
			R (7)										
r2/q2	0.682	0.506	0.591	r2/q2	0.784	0.655							
rms	0.117	0.148	0.101	rms	0.096	0.121							
N	392	98	392	N	490	490							
max	0.47	0.42	0.47	max	0.47	0.47							
min	−0.77	−0.79	−0.77	min	−0.79	−0.79							

[a] π: (1) full descriptor set consisting of 18 E-state descriptors plus the total charge of the substituent; (2) 18 E-state descriptors only. MR: (1) full descriptor set consisting of 18 E-state descriptors plus the molecular weight of the substituent; (2) 18 E-state descriptors only. F: (1) full descriptor set consisting of a total of 23 descriptors including 14 modified KSP descriptors (A) + 5 fourth-shell descriptors (B) + 4 total charges of 4 shells (C); (2) A+B; (3) A; (4) A+C (C: only the total charges of the shells 1−3 were considered). R: (1) full descriptor set consisting of a total of 24 descriptors including 14 modified KSP descriptors (A) + 5 fourth-shell descriptors (B) + 4 total charges of 4 shells (C) + 1 indicator descriptor (D); (2) A+B+C; (3) A+B; (4) A; (5) A+D; (6) A+C+D (C: only the total charges of the shells 1−3 were considered); (7) A+B+D. "Full" stands for a full data set combining both training and test sets.

To understand the significance of various descriptors, a sensitivity analysis was conducted for all four substituent constants. The results are shown in Table 5. The results indicate that using full sets of descriptors yield the best training and test set results. The importance of total charge to π is revealed when comparing the results of π(1) and π-(2). The correlation coefficients (r^2) drop significantly from 0.839, 0.837, and 0.810 to 0.678, 0.733, and 0.597 for

training, test, and cross-validation sets, respectively. The same trend was also observed for a full data set (i.e., when both training and test sets were combined). For MR, molecular weight plays an important role, as demonstrated by the significantly poorer statistics in both test and cross-validation sets when molecular weight was removed.

Comparing the results of F(1)−F(4) provides hints as to which descriptors are more essential. It is interesting to note that F(4) is marginally worse than F(1), indicating that the fourth-shell atoms are probably too distant from the attachment atom to affect F in any significant manner. On the other hand, the total charge of each shell is important, since removing it from the descriptor set results in a substantial decline of correlation coefficients in all runs, as can be seen from the results of F(3) and F(4).

The result of the sensitivity study for R is also quite interesting. R(1) performed significantly better than all the other descriptor sets. This implies that every descriptor in R(1) is important in determining R. Specifically, comparing R(1) with R(6) reveals that the fourth-shell atoms still bear on the resonance effect despite a long through-bond distance effect. In contrast, field effect does not carry to this extent.

With this set of substituent constant prediction models, the next steps include application to aspects of the drug discovery process such as (1) bioisosteric replacement based on clustering of similar Hansch parameters; (2) combinatorial chemistry library design by surveying representative substituents in the Hansch parameter space; (3) SAR exploration; and (4) generating parameters to encode unnatural amino acids for peptide QSAR applications. Examples of QSAR applications will be discussed in the companion paper.[28]

SUMMARY

Neural network QSPR studies were conducted to correlate four substituent constants (π, MR, F and R) with two different molecular descriptor sets for hundreds of chemically diverse substituents. For π and MR, E-state descriptors were used for correlation, while for F and R, the modified KSP descriptors were adopted. Both descriptor sets were two-dimensional in nature and were thus simple to implement, fast to calculate, and conformation-independent. Both QSPR models demonstrated good predictivity in all our test sets and in leave-one-out cross-validation procedures.

ACKNOWLEDGMENT

T.-L.C. is grateful to Hoffmann-La Roche for funding her postdoctoral research. We are grateful to Dr. Hongmao Sun for insightful discussion on this work and Drs. Robert Goodnow and Nora McDonald for helpful comments on this manuscript.

Supporting Information Available: The Hansch substituent constants (π, MR, F and R) of 764 substituents. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R.; Fujita, T.; Peyton, P.; Maloney, P. et al. The correlation of biological activity of plant growth regulators and chloromycetin deviates with Hammett constants and partition coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(2) Kumar, R.; Singh, P. Quantitative structure−activity relationship study of iodinated analogues of trimetoquinol as highly potent beta 2-adrenoceptor ligands. *Indian J. Biochem. Biophys.* **1998**, *35*, 390−392.

(3) Augelli-Szafran, C. E.; Horwell, D. C.; Kneen, C.; Ortwine, D. F.; Pritchard, M. C. et al. Cholecystokinin B antagonists. Synthesis and quantitative structure−activity relationships of a series of C-terminal analogues of CI-988. *Bioorg. Med. Chem.* **1996**, *4*, 1733−1745.

(4) Luco, J. M.; Ferretti, F. H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392−401.

(5) Gupta, S. P.; Saha, R. N.; Mulchandani, V. Quantitative structure−activity relationship studies on benzodiazepine receptor binding: recognition of active sites in receptor and modelling of interaction. *J. Mol. Recognit.* **1992**, *5*, 75−80.

(6) Compadre, C. M.; Hansch, C.; Klein, T. E.; Petridou-Fischer, J.; Selassie, C. D. et al. Separation of electronic and hydrophobic effects for the papain hydrolysis of substituted N-benzoylglycine esters. *Biochim. Biophys. Acta* **1991**, *1079*, 43−52.

(7) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995; 1−2, 11−14, 103−105.

(8) Hansch, C.; Leo, A.; Taft, R. W. A survey of Hammett substituent constants and resonance and field parameters. *Chem. Rev.* **1991**, *91*, 165−195.

(9) Hansch, C.; Leo, A.; Hoekman, D. H. *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; American Chemical Society: Washington, DC, 1995; pp 219−304.

(10) Sullivan, J. J.; Jones, A. D.; Tanji, K. K. QSAR treatment of electronic substituent effects using frontier orbital theory and topological parameters. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1113−1127.

(11) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Neural network classification of inductive and resonance effects of substituents. *J. Am. Chem. Soc.* **1993**, *115*, 1495−1500.

(12) Ren, S.; Wang, R.; Komatsu, K.; Bonaz-Krause, P.; Zyrianov, Y. et al. Synthesis, biological evaluation, and quantitative structure−activity relationship analysis of new Schiff bases of hydroxysemicarbazide as potential antitumor agents. *J. Med. Chem.* **2002**, *45*, 410−419.

(13) Wildman, S. A.; Crippen, G. M. Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

(14) Mannhold, R.; Rekker, R. F. The hydrophobic fragmental constant approach for calculating log P in octanol/water and aliphatic hydrocarbon/water systems. *Perspect. Drug. Discov. Design.* **2000**, *18*, 1−18.

(15) Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. Prediction of *n*-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1407−1421.

(16) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure−activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 21−35.

(17) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three-dimensional structure directed quantitative structure−activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(18) Amat, L.; Carbo-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999**, *42*, 5169−5180.

(19) Vaz, R. J. Use of electron densities in comparative molecular field analysis (CoMFA): A quantitative structure activity relationship (QSAR) for electronic effects of groups. *Quant. Struct-Act. Relat.* **1997**, *16*, 303−308.

(20) Kier, L. B.; Hall, L. H. Estimation of substituent group electronic influence from molecular connectivity delta values. *Quant. Struct-Act. Relat.* **1983**, *16*, 163−167.

(21) Ertl, P. Simple quantum chemical parameters as an alternative to the Hammett sigma constants in QSAR studies. *Quant. Struct-Act. Relat.* **1997**, *16*, 377−382.

(22) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.

(23) Huuskonen, J. Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 491−497.

(24) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(25) Huuskonen, J. J.; Livingstone, D. J.; Tetko, I. I. Neural network modeling for estimation of partition coefficient based on atom-type electrotopological state indices. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 947−955.

(26) Hall, L. H.; Kier, L. B. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.
(27) www.daylight.com.
(28) Chiu, T. L.; So, S. S. Development of neural network QSPR nodels for Hansch substituent constants. 2. Applications in QSAR studies of HIV-1 reverse transcriptase and dihydrofolate reductase inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 154−160.

CI030293Q