J. Org. Chem. **2001,** *66,* 4036−4040

A QSRR Treatment of Solvent Effects on the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylates Employing Molecular Descriptors

Alan R. Katritzky,* Subbu Perumal, and Ruslan Petrukhin

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 117200, Gainesville, Florida 32611-7200

katritzky@chem.ufl.edu

Received August 4, 2000

A quantitative structure−reactivity relationship (QSRR) study of the decarboxylation rates of 6-nitrobenzisoxazole-3-carboxylic acid (Kemp, D. S.; Paul, K. G. *J. Am. Chem. Soc.* **1975**, *97*, 7305). employing the CODESSA program correlates the effect of 24 solvents with theoretical descriptors to provide a straightforward interpretation of these solvent effects in terms of molecular parameters.

Introduction

Kemp and Paul[1] first investigated solvent effects on the rates of decarboxylation of tetramethylguanidinium 6-nitrobenzisoxazole-3-carboxylate (**1**) to afford salicylonitriles (**2**). The reaction rates depend greatly on the nature of the 24 pure solvents employed, spanning up to 8 orders of magnitude from the protic solvent, water to the dipolar aprotic solvent, hexamethylphosphoramide (Table 1). The log *k* values of the decarboxylation of **1** gave a scatter plot vs the Dimroth parameter, E_T. A plot of the log *k* values of the molar transition energy, *E*, of the long-wavelength electronic absorption maximum of the tetramethylguanidinium (TMG) salt of 2-cyano-5-nitrophenol gave three straight lines dividing the solvents into three groups, although no common structural feature characterized the solvents in each of these classes.

Subsequently, four other groups employed different solvent scales to analyze these rate data. Grate et al.[2] used five Taft and Kamlet experimentally determined parameters[3] to obtain eq 1 for 20 solvents (for the remaining four solvents, these parameters were unavailable at that time).

$$\log k = (5.45 \pm 0.89)\pi^* - (1.46 \pm 0.45)\delta - (3.03 \pm 0.65)\alpha + (1.80 \pm 0.66)\beta - (1.06 \pm 0.25)\delta^2{}_{\mathrm{H}} - (2.97 \pm 0.52) \quad (1)$$

$$R^2 = 0.920,\ \mathrm{SD} = 0.58,\ n = 20$$

Famini and Wilson[4] applied seven theoretical parameters. Ferris and Drago[5] divided the solvents into three sets to derive correlation equations.

$$\log k = (0.26 \pm 1.23)V_{\mathrm{mc}} - (25.07 \pm 54.81)\pi_1 - (48.09 \pm 31.37)\epsilon_{\mathrm{B}} + (4.87 \pm 2.07)q_- + (27.21 \pm 27.87)\epsilon_{\mathrm{A}} - (18.14 \pm 6.48)q_+ - (0.95 \pm 1.06)\delta^2{}_{\mathrm{H}} - (9.08 \pm 4.44) \quad (2)$$

$$R^2 = 0.767,\ \mathrm{SD} = 1.27,\ n = 24$$

Recently, all this work has been critically overviewed and discussed by Catalan et al.[6] These workers obtained correlation equations 2−4 for 24 solvents employing the solvent scales of Famini and Wilson,[4] Taft and Kamlet,[3] and Drago et al.,[5] respectively. Using their own experimentally determined solvent scales, Catalan et al.[6] derived eq 5 for 24 pure solvents and eq 6 for 60 pure and mixed solvents.

$$\log k = (5.23 \pm 0.95)\pi^* - (1.58 \pm 0.62)\delta - (3.77 \pm 0.65)\alpha + (1.36 \pm 0.74)\beta - (0.84 \pm 0.27)\delta^2{}_{H} - (2.80 \pm 0.56) \quad (3)$$

$$R^2 = 0.920,\ \mathrm{SD} = 0.70,\ n = 24$$

$$\log k = (0.71 \pm 1.74)S' + (6.07 \pm 1.96)E - (2.56 \pm 2.44)C - (6.75 \pm 4.58) \quad (4)$$

$$R^2 = 0.421,\ \mathrm{SD} = 4.42,\ n = 24$$

$$\log k = (10.37 \pm 1.47)SPP - (5.93 \pm 0.58)SA + (2.59 \pm 0.74)SB - (9.74 \pm 1.17) \quad (5)$$

$$R^2 = 0.904,\ \mathrm{SD} = 0.73,\ n = 24$$

$$\log k = (10.03 \pm 1.05)SPP - (5.73 \pm 0.40)SA + (2.41 \pm 0.49)SB - (9.58 \pm 0.83) \quad (6)$$

$$R^2 = 0.870,\ \mathrm{SD} = 0.60,\ n = 60$$

Previous QSPR/QSAR studies of various physicochemical properties and biological activities using the CODES-

---

[1] Kemp, D. S.; Paul, K. G. *J. Am. Chem. Soc.* **1975**, *97*, 7305.

[2] Grate, J. W.; McGill, R. A.; Hilvert, D. *J. Am. Chem. Soc.* **1993**, *115*, 8577.

[3] (a) Kamlet, M. J.; Taft, R. W. *J. Am. Chem. Soc.* **1976**, *98*, 377. (b) Taft, R. W.; Kamlet, M. J. *J. Am. Chem. Soc.* **1976**, *98*, 2886. (c) Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877.

[4] Famini, G. R.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1994**, 1641.

[5] Ferris, D. C.; Drago, R. S. *J. Am. Chem. Soc.* **1994**, *116*, 7509.

[6] Catalan, J.; Diaz, C.; Garcia-Blanco, F. *J. Org. Chem.* **2000**, *65*, 3409.

10.1021/jo0011843 CCC: $20.00

Table 1. log *k* Values for the Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylate Ions in Different Solvents

						log *k* (calcd)[a]					
no.	solvent name	log *k* (exptl)	^{0}SIC	SE_{OK}	HASA	P_1	ϵ_1	P_2	ϵ_2	P_3	ϵ_3
1	water[b]	−5.13	1.74	0.19	57.31	−3.72	1.41	−4.76	0.37	−5.43	−0.30
2	methanol[b]	−3.60	2.91	0.08	41.79	−2.51	1.09	−3.62	−0.02	−3.85	−0.25
3	formamide[b]	−3.13	4.16	0.30	68.27	−1.20	1.93	−3.52	−0.39	−3.22	−0.09
4	chloroform	−3.09	2.95	0.01	0.00	−2.46	0.63	−1.56	1.53	−2.20	0.89
5	ethanol[b]	−3.00	3.48	0.05	31.04	−1.91	1.09	−2.73	0.27	−2.66	0.34
6	carbon tetrachloride	−2.82	1.55	0.00	0.00	−3.91	−1.09	−3.07	−0.25	−3.11	−0.29
7	benzene	−2.32	3.35	0.00	0.00	−2.05	0.27	−1.64	0.68	−2.06	0.26
8	*N*-methylformamide[b]	−2.09	5.34	0.24	58.13	0.02	2.11	−2.13	−0.04	−1.70	0.39
9	dimethoxymethane	−1.44	4.69	0.11	0.00	−0.65	0.79	−0.25	1.19	0.05	1.49
10	dioxane	−1.39	5.07	0.00	0.00	−0.26	1.13	−0.04	1.35	−1.06	0.33
11	dichloromethane	−1.33	3.28	0.03	0.00	−2.12	−0.79	−1.21	0.12	−1.88	−0.55
12	diethyl ether	−1.05	4.45	0.02	0.00	−0.90	0.15	−0.51	0.54	−0.65	0.40
13	nitromethane	−0.24	4.59	0.28	0.00	−0.75	−0.51	−0.12	0.12	1.10	1.34
14	benzonitrile	0.40	5.67	0.11	0.00	0.37	−0.03	0.52	0.12	0.83	0.43
15	acetonitrile	0.46	4.16	0.20	0.00	−1.20	−1.66	−0.45	−0.91	0.15	−0.31
16	tetrahydrofuran	0.60	4.35	0.05	0.00	−1.00	−1.60	−0.71	−1.31	−0.41	−1.01
17	diglyme	0.70	6.74	0.02	0.00	1.47	0.77	1.59	0.89	0.75	0.05
18	dimethyl sulfoxide	1.00	4.73	0.20	0.00	−0.61	−1.61	−0.15	−1.15	0.45	−0.55
19	acetone	1.38	4.73	0.15	0.00	−0.61	−1.99	−0.15	−1.53	0.00	−1.38
20	dimethylformamide	1.56	5.96	0.19	0.00	0.67	−0.89	0.98	−0.58	1.72	0.16
21	tetramethylenesulfone	1.81	6.30	0.15	0.00	1.02	−0.79	1.10	−0.71	1.54	−0.27
22	dimethylacetamide	2.20	6.48	0.16	0.00	1.21	−0.99	1.42	−0.78	1.52	−0.68
23	*N*-methylpyrrolidone	2.40	6.87	0.13	0.00	1.61	−0.79	1.62	−0.78	2.26	−0.14
24	hexamethylphosphoramide	2.80	9.38	0.13	0.00	4.22	1.42	4.04	1.24	2.53	−0.27

[a] These values were calculated using correlations from Table 2. P_1 is the predicted value from correlation 1; ϵ_1 is the absolute error. [b] Protic solvents.

SA program[7] have led to enhanced correlations. Significantly, analysis of 40 different solvent scales by multilinear regression analysis[8] and by principal component analysis[9] facilitated comparisons between the different scales as well as revealing characteristic features of the individual solvents.

The process of decarboxylation of the tetramethylguanidinium salt of 6-nitrobenzisoxazole-3-carboxylic acid and its large solvent effect is an important biochemical model.[6] We therefore applied CODESSA to this problem and present the results in this paper.

Methodology

Optimization of the geometrical structures of the solvent molecules was performed using the AM1 method[10] within the MOPAC program.[11] The geometry and other information from the output of the quantum chemical computations were fed into the CODESSA program,[12] and the descriptors were calculated. All these descriptors were obtained solely from the molecular structures without requiring any experimental data. Hence, the rate data in all the 24 solvents could be correlated with the molecular descriptors in the present study. Both the heuristic (HM)[12] and the best multilinear correlation regression (BMLR)[12] methods were used to arrive at the optimum correlations. These methods employ different approaches in the discrete local optimization of the correlation coefficient in the descriptor permutation space. As they are local optimization methods, there is no guarantee that the global optimum on the discrete descriptor space will be found by them. Although both methods use heuristics for eliminating the descriptors with low probability to be involved in the final multiparameter correlation, the approaches to this elimination are completely different.

The best multilinear regression (BMLR) method[12] searches for the multiparameter regression with the maximum predicting ability using the following strategy:

Step 1. Setting up initial values: the upper limit of the square of the linear correlation coefficient $R^2{}_{\min}$, for two descriptor scales to be considered orthogonal (default value $R^2{}_{\min} = 0.1$); the lower limit of the square of the linear correlation coefficient $R_{nc}{}^2$, for two descriptor scales to be considered noncollinear (default value $R_{nc}{}^2 = 0.65$); the maximum number of best correlations to be considered in the search of the best correlations with one descriptor scale added (default value $N_c = 400$); the probability level for the Fisher criterion F_{Sel} (default value 0.95).

Step 2. All orthogonal pairs of descriptors *i* and *j* (with $R_{ij}{}^2 > R^2{}_{\min}$) are found in a given data set.

Step 3. The property analyzed is treated by using the two-parameter regressions with the pairs of descriptors, obtained in step 2. The N_c pairs with highest regression correlation coefficients are chosen for performing the higher-order regression treatments.

Step 4. For each descriptor set, obtained in the previous step, a noncollinear descriptor scale, *k* (with $R_{ik}{}^2 < R_{nc}{}^2$ and $R_{kj}{}^2 < R_{nc}{}^2$), is added, and the respective $(n + 1)$-parameter regression treatment is performed. If the Fisher criterion at a 95% probability level, *F*, is smaller than that for the best *n*-parameter correlation from the previous step, the latter is chosen as the final result and the program proceeds to the output section (step 5). Otherwise, the N_c descriptor sets with highest regression

---

[7] (a) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, 279. (b) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551. (c) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.

[8] Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 684.

[9] Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 692.

[10] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

[11] Stewart, J. J. P. *MOPAC 7.0 for Unix*, Quantum Chemistry Program Exchange, Project 688.

[12] Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference Manual*; *Version 2*; University of Florida: Gainesville, 1996.

Table 2. Results of Correlation Analysis of the Rate of Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylic Acids in 24 Pure Solvents

P no.	R^2	F	SD	s^2	R^2_{CV}		B	ΔB	t	name of descriptor
1	0.688	48.53	1.25	1.570	0.63	0	−5.5263	0.7469	−7.34	intercept
						1	1.0389	0.1491	6.97	^{0}SIC
2	0.843	56.57	0.91	0.826	0.80	0	−4.7384	0.6535	−7.25	intercept
						1	0.92654	0.12039	7.70	^{0}BIC
						2	−0.04497	0.008860	−5.06	HASA
3	0.909	66.21	0.71	0.507	0.87	0	−4.0132	0.4360	−9.21	intercept
						1	0.58251	0.07632	7.36	^{1}SIC
						2	−0.07093	0.007761	−9.14	HASA
						3	8.5412	1.8779	4.55	SE_{OK}
4	0.944	80.43	0.57	0.325	0.92	0	−3.4041	0.3330	−10.22	intercept
						1	0.51304	0.09693	5.29	$CSA2_H$
						2	8.4600	2.0508	4.13	HDSA
						3	−0.04946	0.004865	−10.17	HBSA
						4	0.56704	0.10173	5.57	D

correlation coefficients are chosen for this step repeated with incremented value of *n*.

Step 5. The final result, with the maximum value of the Fisher criterion and with the highest cross-validated correlation coefficient, is chosen for use in the following prediction part of the program. The following results for the best correlation of each rank obtained are saved.

The heuristic method (HM) for descriptor selection, by contrast, proceeds with a preselection of descriptors by eliminating (step by step) the following descriptors: those that do not match any of the following criteria: (1) Fisher *F*-criterion greater than one unit; (2) R^2 value less than a value defined at the start (by default 0.01); (3) Student's *t*-criterion less than that defined (by default 0.1); and those having a higher squared intercorrelation coefficient than a predetermined level (usually 0.8). Descriptors with higher R^2 with reference to the property are retained.

This is followed by listing the remaining descriptors in decreasing order of the correlation coefficients when used in global search for two-parameter correlations. Each significant two-parameter correlation by *F*-criterion is recursively expanded to an *n*-parameter correlation until the normalized *F*-criterion remains greater than the startup value. The top *N* correlations by R^2 as well as *F*-criterion are saved. The HM usually produces correlations 2−5 times faster than BMLR, with quality comparable to that obtained with the latter. The rapidity of calculations from the HM renders it the first method of choice in practical research.

Results and Discussion

The log *k* values for the decarboxylation of tetramethylguanidinium salt of 6-nitrobenzisoxazole-3-carboxylic acid (**1**) in 24 solvents measured by Kemp and Paul[1] and listed in Table 1 were correlated with molecular descriptors of the solvent molecules, as discussed above, to arrive at the optimum one- and multiparameter correlations listed in Table 2.

With the 24 data points available in the present study, one-, two-, three-, or four-parameter correlations are statistically significant according to the criterion that the ratio of data points to descriptors should be greater than 5:1, and the results are furnished in Table 2. We report several correlations (from one- to four-parameter) in this study with a view to illustrate the ability of the CODESSA program to select a variety of significant correlations using the HM and BMLR methods. The correlations include descriptors of three types: (i) information context descriptors (^{0}BIC, ^{0}SIC, ^{1}SIC) that measure the structural diversity, (ii) the polarity descriptor (SE_{OK}, etc.), and (iii) hydrogen-bonding descriptors (HASA1, HASA1/TMSA, etc.). The descriptors of each type are highly correlated with one another but are not highly correlated between the groups (see Table S1 in the Supporting Information). Table S2 gives the intercorrelation of the descriptors within each single correlation. The intercorrelation became significant at the four-parameter level: we therefore limit our detailed discussion to the equations with up to three descriptors.

Although the inclusion of different descriptors in different multiparameter correlations may be viewed as an apparent shortcoming of CODESSA program, deeper analysis indicates that this is not so. In the present study, the descriptors in the equations fall into three classes, and within each class the descriptors intercorrelate highly with each other. This suggests that the selection of a different descriptor belonging to each class of descriptors in different multiparameter correlations is mainly due to the availability of several sets of interrelated descriptors (from CODESSA calculations) having similar chemical significance. Hence, the different descriptors in each class arrived at in the different correlation equations point to the common factors underlying the solvent effect. The values of representative descriptors of each type (^{0}SIC, SE_{OK}, and HASA) are depicted in Table 1.

The two- and the three-parameter correlations of the Table 2 each involve one topological parameter,[13] the bonding information content (order 0), ^{0}BIC and structural information content (order 1), ^{1}SIC, respectively. This indicates that the branching and connectivity of the solvent molecules have an important bearing on the interactions of the solvent with the substrate and/or the transition state in the decarboxylation process. These correlations also have as a common descriptor the hydrogen acceptor accessible surface area, HASA. The negative coefficient for this descriptor implies that enhanced hydrogen bonding interactions of the carboxylate with the solvent decreases the rate of the reaction. This is understandable in view of the fact that the hydrogen bonding would stabilize the ground state viz., the carboxylate ion and impede the decarboxylation as concluded in the previous studies.[1,2]

The three-parameter correlation discloses a dominant contribution of the image of the Onsager−Kirkwood

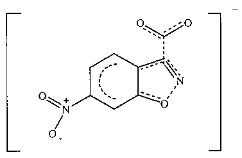
---

[13] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

Table 3. Results of Correlation Analysis of the Rate of Decarboxylation of 6-Nitrobenzisoxazole-3-carboxylic Acids in 19 Pure Nonprotic Solvents

P no.	R^2	F	SD	s^2	R^2_{CV}		B	ΔB	t	name of descriptor
1	0.707	41.10	1.00	0.997	0.63	0	−4.6178	0.7686	−6.00	intercept
						1	0.90528	0.14120	6.41	^{0}BIC
2	0.852	46.07	0.73	0.536	0.81	0	−2.2838	0.2998	−7.62	intercept
						1	−4.6959	0.5876	−7.99	q_{min}
						2	2.3910	0.3661	6.53	N_{HA}
3	0.929	64.98	0.53	0.276	0.89	0	−6.9288	1.1733	−5.91	intercept
						1	0.48537	0.05256	9.23	HACA1
						2	−5.3169	0.4576	−11.62	q_{min}
						3	9.7983	2.4024	4.09	V_{Mol}/V_{XYZ}

Chart 1. Transition State



(3)

solvation energy (SE_{OK}). This descriptor is defined as

$$SE_{OK} = \frac{D^2}{MW}$$

where D is dipole moment of the molecule calculated on the basis of AM1 Mulliken charges in our case and MW is the molecular weight. Hence, this descriptor is a measure of the polarity of the solvent. The positive coefficient for SE_{OK} indicates that with an increase in the value of this descriptor the energy of activation for the decarboxylation process diminishes. This could presumably be ascribed to the diminution in the ion pairing of the reactant salt with enhanced polarity of the solvent. This is also in accord with the conclusions reached in the earlier studies.[5] Each of the descriptors in the three-parameter correlation (Table 2, correlation 3) is among those frequently encountered in the theoretical models evolved for the different solvent scales in our previous study.[8] This leads to the conclusion that the selection of the descriptors by CODESSA is soundly based and that the various factors that contribute to the solvent effect involved in the present study are among the factors that comprise various solvent scales. Hence, the CODESSA program now used for the first time in this type of study should find utility in future investigations of solvent effects on other bio-model reactions. It is pertinent to note that CODESSA correlations are straightforward as the molecular descriptors for all solvents can be calculated rapidly without (i) prejudging which solvent scale will be best suited for segregating the contributions of the different factors involved in a new study and (ii) correlating the data with different solvent scales in a trial and error manner to obtain the best description of the nature and composition of the factors that are involved in a particular reaction.

The descriptors in the four parameter correlation, charged surface area CSA^2_H, hydrogen donor accessible surface area HDSA, hydrogen-bonding surface area HBSA, total dipole of the molecule D, and hydrogen acceptor part of the total molecular solvent accessible surface area HASA-2/TMSA are in consonance with the transition state (**3**, Chart 1). The contribution of the topological and charge related parameters in the above correlations may be linked to ion-pair formation and the associated equilibrium, besides their other influences in stabilizing/destabilizing the reactant/transition state. For instance, the descriptors hydrogen donor accessible surface area (HDSA) and hydrogen bonding surface area (HBSA), respectively, reflect the solvent hydrogen bond basicity and acidity. The positive and negative descriptor coefficients of these descriptors, HDSA and HBSA, are in accord with the effects of solvent hydrogen bond basicity and acidity on the solvent effect concluded in previous studies.[1,2]

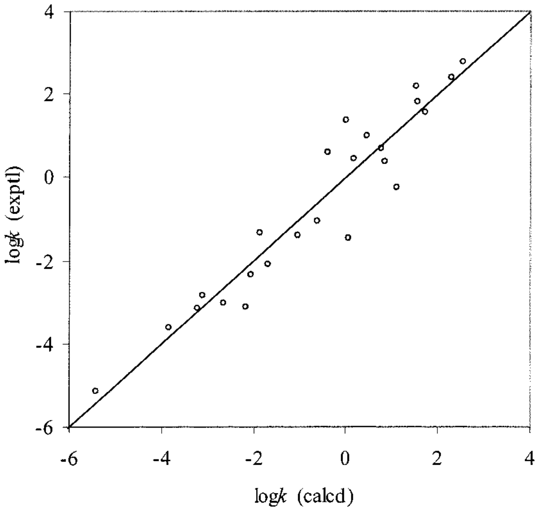


Figure 1. Plot of log k (exptl) versus log k (calcd) using the three-parameter correlation.

An increase in the bulk of the groups attached to solvent molecules and a decrease in the polarity of the solvent would facilitate ion-pair formation. The plot of log k (exptl) vs log k (calcd) employing the three-parameter correlation (Table 2) is depicted in Figure 1.

Excluding the protic solvents (water, methanol, ethanol, formamide, and N-methylformamide) does not significantly improve the correlations (Table 3), although it leads to the involvement of different descriptors in the correlations. However, the simple topological descriptor, ^{1}SIC alone accurately describes the rate variation in the case of the five protic solvents in eq 7. The structural information content topological index is based in the Shannon information theory and is defined as

$$^k\mathrm{SIC} = \frac{^k\mathrm{IC}}{\log_2 n}$$

$$^k\mathrm{IC} = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

where n_i is a number of atoms in the ith class, n is the total number of atoms in the molecule, and k is the number of atomic layers in the coordination sphere around a given atom that are accounted for (in our case, $k = 0, 1$).

This shows clearly that among these solvents the rate variation in these solvents is almost wholly influenced by the structure of the solvent molecules, the other factors remaining almost constant.

$$\log k = (0.52965 \pm 0.03366)\,{}^{1}\mathrm{SIC} - (5.9284 \pm 0.1732) \quad (7)$$

$$R^2 = 0.998,\ F = 247.57,\ s^2 = 0.0199,\ \mathrm{SD} = 0.14,\ n = 5$$

A comparison of the present results with those reported previously shows that (i) our three-parameter correlation given in Table 2 describes the solvent effect better than the seven-parameter equation (2) and also eq 4 obtained by fitting the rate data with the solvent scales of Drago et al., (ii) our three-parameter correlation is almost as successful as the five-parameter equations (1) and (3) and comparable to that of the three-parameter equation (5) of Catalan et al.[6]

Conclusions

Catalan et al.'s[6] successful treatment of mixed and pure solvents in a single scale is presently beyond the scope of CODESSA. However, for pure solvents, our treatment utilizing solely theoretical descriptors derived entirely from molecular structure of the solvents, has the significant advantage that it can be applicable to any solvent in contrast to the previous treatments which require the availability of experimentally determined descriptors. The present work provides a example of the utility of the QSPR treatment involving descriptors derived solely from chemical structure to a model reaction of considerable biological significance.

Supporting Information Available: Tables S1 and S2, "Descriptor intercorrelations (R^2 values)". This material is available free of charge via the Internet at http://pubs.acs.org.

JO0011843