Prediction of Critical Micelle Concentration Using a Quantitative Structure−Property Relationship Approach. 1. Nonionic Surfactants

Paul D. T. Huibers,[†,‡] Victor S. Lobanov,[§,||] Alan R. Katritzky,*[,§] Dinesh O. Shah,[†,⊥] and Mati Karelson[▽,#]

Center for Surface Science and Engineering, Department of Chemical Engineering, University of Florida, P.O. Box 116005, Gainesville, Florida 32611-6005, Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 117200, Gainesville, Florida 32611-7200, and Department of Chemistry, University of Tartu, 2 Jakobi Str., Tartu, EE 2400, Estonia

Received July 14, 1995. In Final Form: November 28, 1995[⊗]

A quantitative structure−property relationship study was performed on the critical micelle concentration (cmc) of nonionic surfactants using the CODESSA program. A known correlation between the logarithm of the cmc and counts of linear alkane carbon atoms and ethoxy groups in linear alkyl ethoxylates was improved ($R^2 = 0.996$) by adding cross terms of these molecular descriptors. A general three-parameter structure−property relationship was developed for a diverse set of 77 nonionic surfactants ($R^2 = 0.984$) employing topological descriptors calculated for the hydrophobic fragment of the surfactant molecule. The three descriptors represent contributions from the size of the hydrophobic group, the size of the hydrophilic group, and the structural complexity of the hydrophobic group.

Introduction

It is well-known that critical micelle concentration (cmc) depends on molecular structure. Several empirical relationships have been reported in the literature relating structural features to the cmc of homologous compounds. Accurate cmc predictions for linear alkyl ethoxylates can be made using parameters such as the alkane carbon number and the ethylene oxide number.[1−5] While the advantage of such empirical predictions is their simplicity of calculation, these relationships are not suitable for a general estimation of cmc for compounds other than the linear alkyl or alkylphenol ethoxylates.

We expect that more general structure−cmc relationships can be developed through the systematic quantitative structure−property relationships (QSPR) approach, using large databases of molecular descriptors. Although a large body of knowledge concerning the prediction of physical properties for organic compounds has been developed over the last two decades with the help of QSPR,[6−12] there is little information available for the prediction of solution properties of surface active compounds.

In recent years, a general QSPR prediction methodology has been developed, coded into software (CODESSA, comprehensive descriptors for structural and statistical analysis),[13] and successfully employed for the prediction of a variety of physical properties of compounds.[14,15] This methodology combines different ways of quantifying the structural information about the molecule with advanced statistical analyses for the establishment of molecular structure−property (activity) relationships. All structural information is encoded into a large number of quantitative descriptors, which reflect the structural, topological, geometrical, electrostatic, and molecular-orbital characteristics of a molecule. In the present paper, the CODESSA program is used for the calculation of descriptors and for statistical analysis, as applied to the prediction of the cmc of nonionic surfactants, given only the *molecular structure* of the surfactants.

Background

The cmc is the concentration at which micelles first appear in the solution. Below this critical value, additional surfactant added to solution remains in monomeric form, and above this, essentially all additional surfactant forms micelles. This transition from premicellar to micellar solutions at the cmc occurs over a narrow range of concentration.

The cmc is an extremely useful quantity, as it captures the surface and interfacial activity of the surfactant molecules in solution. The ability of surfactants to reduce surface or interfacial tension is expected to be directly

* To whom correspondence should be addressed. Email: katritzk@pine.circa.ufl.edu.
† Department of Chemical Engineering, University of Florida.
‡ Email: huibpd@che.ufl.edu.
§ Department of Chemistry, University of Florida.
|| Email: victor@ufark2.chem.ufl.edu.
⊥ Email: shah@che.ufl.edu.
▽ University of Tartu.
Email: mati@chem.ut.ee.
⊗ Abstract published in *Advance ACS Abstracts,* February 1, 1996.

(1) Rosen, M. J. *J. Colloid Interface Sci.* **1976**, *56*, 320.
(2) Meguro, K.; Takasawa, Y.; Kawahashi, N.; Tabata, Y.; Ueno, M. *J. Colloid Interface Sci.* **1981**, *83*, 50.
(3) Becher, P. *J. Dispersion Sci. Technol.* **1984**, *5*, 81.
(4) Ravey, J. C.; Gherbi, A.; Stebe, M. J. *Prog. Colloid Polym. Sci.* **1988**, *76*, 234.
(5) Rosen, M. J. *Surfactants and Interfacial Phenomena,* 2nd ed.; Wiley: New York, 1989.
(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(7) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.
(8) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. *Dokl. Akad. Nauk SSSR* **1987**, *296*, 883.
(9) Stankevich, M. I.; Stankevich, I. V.; Zefirov, N. S. *Russ. Chem. Rev.* **1988**, *57*, 191.
(10) Rouvray, D. H. *J. Mol. Struct.* **1989**, *185*, 187.
(11) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(12) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947.
(13) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 1.1 Reference Manual, 1994.
(14) Murugan, R.; Grendze, M. P.; Toomey, J. E.; Katritzky, A. R.; Karelson, M.; Lobanov, V. S.; Rachwal, P. *CHEMTECH* **1994**, *24*, 17.
(15) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. *Anal. Chem.* **1994**, *66*, 1799.

0743-7463/96/2412-1462$12.00/0

Table 1. Empirical Relationships for Nonionic Surfactant cmc

surfactant series	coefficients A	coefficients B	T (°C)	ref
Change in the Alkane Carbon Number				
C_nE_6	1.8	0.49	25	[1] Rosen, 1976
C_nE_8	2.1	0.51	15	[2] Meguro, 1981
C_nE_8	1.8	0.50	25	[2] Meguro, 1981
C_nE_8	1.6	0.48	40	[2] Meguro, 1981
Change in the Ethoxylate Number				
$C_{12}E_n$	−4.47	−0.056	27	[4] Ravey, 1988

Table 2. Coefficients for Multiple Parameter Fits of Nonionic Surfactant cmc Values, for the Linear Alkyl Ethoxylates at 300 K

coefficients A	coefficients B	coefficients C	coefficients D	T (°C)	ref
1.46	−0.50	0.060		25	[3] Becher, 1984
1.77	−0.52	0.032	0.002	27	[4] Ravey, 1988

related to the cmc. A low cmc indicates that it is thermodynamically favorable for the hydrophobic domain of the surfactant molecule to leave the aqueous solution, which will result in both an excess concentration at the interface and the formation of micelles. This ability to adsorb at an interface and reduce interfacial tension is of great importance to many processes of technological interest, such as emulsification, foaming, wetting, solubilization, detergency, particle suspensions, and surface coatings. The cmc is also important in thermodynamic studies of micellar solutions, inasmuch as it is related to the free energy of micelle formation.

The importance of the cmc has led to its measurement by many researchers, for a wide range of surfactants, and under different solvent conditions. Experimental determination of cmc can be performed rather accurately, often within a range of uncertainty of just a few percent.[16] Our survey of the literature has resulted in 77 cmc values for homogeneous (monomerically pure) nonionic surfactants in aqueous solution at 25 °C.

Previous Predictions of the cmc for Nonionic Surfactants. Several investigators have developed empirical relationships between the cmc and the structural features of surfactants. However, all of these have been limited to a homologous series of surfactants, such as the linear alkyl ethoxylates. A linear relationship (eq 1) has been found between the logarithm of the cmc and the number of alkane carbon atoms n for homologous series of linear alkyl hexaethoxylates (C_nE_6)[1] and octaethoxylates (C_nE_8)[2] (see Table 1). This relationship is

$$\log_{10} \text{cmc} = A - Bn \quad (1)$$

where A and B are empirical regression coefficients. Similar relationships have been found between the cmc and the number of ethylene oxide residues.[4] For dodecyl polyethoxylates ($C_{12}E_n$), the regression coefficients are given in Table 1.

Empirical fits have been determined for the linear alkyl ethoxylate surfactants in general (C_nE_m), using both the carbon number, n, and the ethylene oxide number, m. Original coefficients have been calculated by Becher.[3] Ravey had improved the correlation equation[4] by including a nonlinear term in the form of the product of the alkane carbon number and the ethylene oxide number, nm. The regression coefficients for eq 2 are presented in Table 2.

$$\log_{10} \text{cmc} = A + Bn + Cm + Dnm \quad (2)$$

A more general method is the molecular thermodynamic approach to predict micellization.[40,41] This has been applied to a number of surfactant properties and successfully predicted cmc for a number of linear alkyl ethoxylate and alkyl glucoside surfactants.

Previous QSPR Studies. To our knowledge, no cmc studies have been performed to date using a general QSPR approach. However, some QSPR studies on other surface phenomena are of interest. The charged partial surface area (CPSA) descriptors have been introduced by Stanton and Jurs to correlate the surface tension of organic molecules.[11] Employing CPSA descriptors, they had developed a six-parameter model ($R^2 = 0.908$, $F = 39$) for a set of 31 organic molecules. This was followed up with a more complete analysis of the surface tension of pure liquid,[17] where 95 alkanes ($R^2 = 0.978$) and 35 alcohols ($R^2 = 0.975$) were regressed to 6 descriptors. A combined set of alkanes, alcohols, and alkyl esters was considered and resulted in a squared correlation coefficient of 0.983 using 10 descriptors. This study demonstrated the usefulness of the CPSA descriptors, along with fractional hydrogen-bonding acceptor area and donor area descriptors, in the prediction of surface properties. In another study,[18] the aqueous solubility of organic compounds was correlated to molecular structure for a set of 91 hydrocarbons ($R^2 = 0.978$), 79 halogenated hydrocarbons ($R^2 = 0.975$), and 92 ethers and alcohols ($R^2 = 0.975$), using 10 descriptors. A squared correlation coefficient of 0.937 was achieved for the combined set of 258 structures. These results demonstrate the value of grouping structures by molecular type, in order to achieve better correlation results. As in the previous study, the CPSA descriptors played an important role.

CODESSA Program. The CODESSA program[13] has been developed at the University of Florida and successfully applied in various QSPR studies of physicochemical properties of organic compounds.[19] In a QSPR treatment on 152 diverse organic compounds, good correlations have been obtained for gas chromatographic retention times ($R^2 = 0.959$) and Dietz response factors ($R^2 = 0.892$).[15] An analysis of five physical properties (boiling point, melting point, flash point, octanol−water partition coefficient, and gas-chromatographic retention index) of substituted pyridines produced good correlations for all properties.[14]

CODESSA is a chemical multipurpose quantitative structure−activity and structure−property (QSAR/QSPR) statistical analysis and prediction program designed for the Microsoft Windows environment. The program can generate a large number (>300) of molecular descriptors on the basis of the constitutional, topological, geometrical, and electronic structure of a molecule. Additionally, a large number of descriptors can be derived from the information produced by semiempirical quantum chemical calculational schemes, such as AM1 or PM3. The statistical analysis techniques available within the CODESSA program include principal component analysis, best multilinear regression analysis, and a heuristic method.

Data and Methodology

Data Set. The data set was chosen to contain only nonionic surfactants with cmc values that are expected to be accurate to better than 10%. The major sources of

(16) Mukerjee, P., Mysels, K. J. NSRDS-NBS 36; U.S. Dept. of Commerce: Washington, DC, 1971.

(17) Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109.

(18) Nelson, T. M.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601.

(19) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.

Table 3. Summary of the Complete Nonionic Surfactant Data Set by Class

nonionic surfactant class	number of molecules
linear alkyl ethoxylates	31
branched alkyl ethoxylates	5
octylphenol ethoxylates	10
alkanediols	5
alkyl mono- and disaccharide ethers and esters	7
ethoxylated alkyl amines and amides	9
fluorinated linear ethoxylates and amides	10
total	77

these values are the compendium by Mukerjee and Mysels for pre-1966 cmc data[16] and Rosen's text for later values.[5] Additional values were obtained from a search of the literature.[4,20−23] The surfactants cover a wide range of structural features (Table 3). Representative structures of these seven nonionic surfactant classes can be seen in Figure 1.

Most nonionic surfactants contain ethylene oxide oligomers in the hydrophilic domain of the molecule. These surfactants often contain a distribution of polyethylene oxide chain lengths rather than a constant number of units. Only *monomerically pure surfactants* were included in the present study. All values used were those measured in purified water, at 25 °C, containing no additional salt or solvent. Some data were not available at precisely 25 °C and thus were calculated using an interpolation of a linear fit of $\log_{10}$ cmc vs $1/T$, using data in the 20−40 °C range.[24]

Structure Entry. The molecular structures were generated using PCMODEL (Serena Software) on an IBM RISC/6000 model 320 workstation. The structures were drawn from scratch, 3D-optimized using an MMX molecular mechanics force field available within PCMODEL, and stored as MOPAC input files.

Descriptor Generation. The molecular descriptors encode all essential structural features of the molecules of interest. Ideally, these descriptors should capture all necessary information on factors which influence the phenomena controlling the physical parameters that we wish to predict. The descriptors generated by CODESSA fall into five categories, which are conventionally defined as the constitutional, topological, geometrical, electrostatic, and quantum chemical descriptors. The first four categories of descriptors are calculated by CODESSA given only the molecular structure. The final category is calculated by CODESSA after MOPAC 6.0 is used to generate the necessary quantum mechanical electron distribution estimates.

Over 400 molecular descriptors programmed into CODESSA were calculated for each surfactant structure.

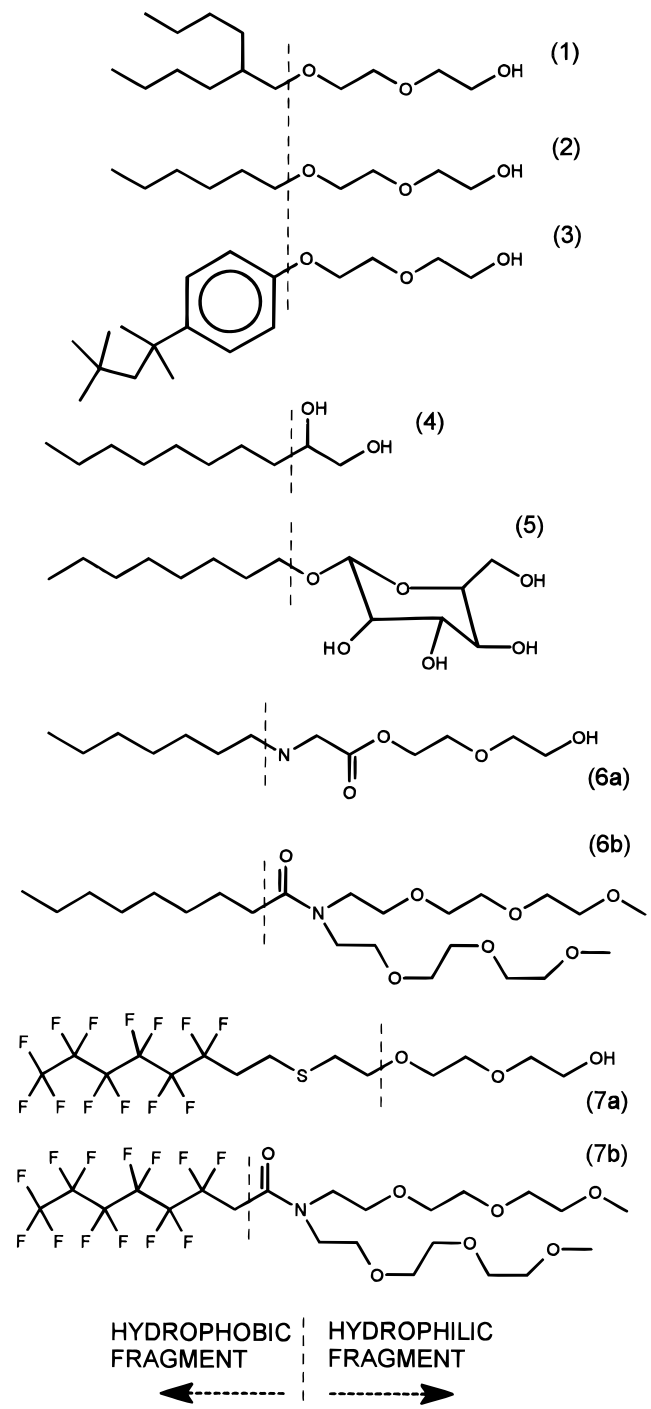

Figure 1. Representative structures of the seven nonionic surfactant classes: (1) branched alkyl ethoxylates, (2) linear alkyl ethoxylates, (3) octylphenyl ethoxylates, (4) alkanediols, (5) alkyl mono- and disaccharides, (6a) ethoxylated alkylamines, (6b) ethoxylated alkylamides, (7a) fluorinated linear alkyl ethoxylates, and (7b) fluorinated ethoxylated amides.

(20) Ravey, J. C.; Stebe, M. J. *Colloids Surf., A* **1994**, *84*, 11.
(21) Matos, L.; Ravey, J. C.; Serratrice, G. *J. Colloid Interface Sci.* **1989**, *128*, 341.
(22) Selve, C.; Ravey, J. C.; Stebe, M. J.; El Moudjahid, C.; Moumni, E. M.; Delpeuch, J. J. *Tetrahedron* **1991**, *47*, 411.
(23) Auvray, X.; Petipas, C.; Anthore, R.; Rico-Lattes, I.; Lattes, A. *Langmuir* **1995**, *11*, 433.
(24) Elworthy, P. H.; Florence, A. T. *Kolloid Z. Z. Polym.* **1964**, *195*, 23.
(25) Stewart, J. J. P. MOPAC 6.0; QCPE No. 455, 1989.
(26) Donbrow, M.; Jacobs, J. *J. Pharm. Pharmacol. Suppl.* **1966**, *18*, 92S.
(27) Corkill, J. M.; Goodman, J. F.; Harrold, S. P. *Trans. Faraday Soc.* **1964**, *60*, 202.
(28) Shinoda, K.; Yamanaka, T.; Kinoshita, K. *J. Phys. Chem.* **1959**, *63*, 648.
(29) Crook, E. H.; Trebbi, G. F.; Fordyce, D. B. *J. Phys. Chem.* **1964**, *68*, 3592.
(30) Hudson, R. A.; Pethica, B. A. In *Chemistry Physics, and Applications of Surface Active Substances*, Overbeek, J. Th.G., Ed.; Proceedings of the International Congress on Surface Active Substances, 4th, Brussels, Sept, 1964; Gordon & Breach: New York, 1964; Vol. 4, p 631.
(31) Rosen, M. J.; Cohen, A. W.; Dahanayake, M.; Hua, X.-Y. *J. Phys. Chem.* **1982**, *86*, 541.
(32) Lange, H. *Kolloid-Z.* **1965**, *201*, 131.
(33) Elworthy, P. H.; MacFarlane, C. B. *J. Pharm. Pharmacol. Suppl.* **1962**, *14*, 100T.
(34) Crook, E. H.; Fordyce, D. B.; Trebbi, G. F. *J. Phys. Chem.* **1963**, *67*, 1987.
(35) Kwan, C.-C.; Rosen, M. J. *J. Phys. Chem* **1980**, *84*, 547.
(36) Shinoda, K.; Yamaguchi, T.; Hori, R. *Bull. Chem. Soc. Jpn.* **1961**, *34*, 237.
(37) Herrington, T. M.; Sahi, S. S. *Colloids Surf.* **1986**, *17*, 103.
(38) Kuwamura, T. In *Structure/Performance Relationships in Surfactants*;Rosen, M. J., Ed.; American Chemical Society: Washington, DC, 1984.

Additionally, some 150 constitutional and topological descriptors were calculated for the hydrophobic and hydrophilic fragments of the surfactant molecules. Several additional descriptors (alkane carbon number, C#, ethoxylate number, EO#, and related terms) were tabulated externally and loaded into CODESSA.

Correlation Analysis. Multiple linear regression analyses of molecular descriptors and the logarithm of the cmc were carried out using the heuristic algorithm available in the CODESSA program.[13] In this algorithm, descriptors are initially screened for insignificance, missing values, and high intercorrelation to limit the number of descriptors considered. The heuristic stepwise addition procedure is then applied to derive the best multiparameter linear correlation equations from this reduced set of descriptors. The final correlation models where chosen on the basis of partial and standard tests of significance and highest correlation coefficients.

The squared correlation coefficient (or coefficient of multiple determination), R^2, is a measure of the fit of the regression model. Correspondingly, it represents the part of the variation in the observed (experimental) data that is explained by the model. The correlation coefficient values closer to 1.0 represent the better fit of the model. The *F*-test reflects the ratio of the variance explained by the model and the variance due to the error in the model (i.e., the variance not explained by the model). High values of the *F*-test indicate that the model is statistically significant.[39] The standard error is measured by the error mean square, s^2, which expresses the variation of the residuals or the variation about the regression line. Thus the standard error measures the model error. If the model is correct, it is an estimate of the error of the data variance, σ^2. The *t*-test measures the statistical significance of the regression coefficients. The higher *t*-test values correspond to the relatively more significant regression coefficients. The resulting linear equations for predicting the physical property of interest are in the form of the following equation

$$P_j = P_0 + \sum_{i=1}^{n} c_i X_{ij} \qquad (3)$$

where P_j is the predicted value of the property for a given compound j, P_0 is the intercept coefficient, c_i are the descriptor coefficients, and X_{ij} are the descriptor values. The residual error terms (ϵ_j) are the differences between the predicted and observed log cmc values. $\Sigma\epsilon_j^2$ is minimized in the multiple linear regression.

Results and Discussion

Prediction of cmc for the Linear Alkyl Ethoxylate Surfactants. The two-parameter empirical model proposed by Becher[3] has been tested by regression analysis of cmc data for the 31 linear alkyl ethoxylates (C_nE_m) in the current database. These two descriptors (the alkane carbon number of the alcohol, C#, and the number of ethylene oxide groups, EO#) resulted in a good correlation with $\log_{10}$ cmc ($R^2 = 0.994$, $F = 2327$, $s^2 = 0.0166$). The coefficient values for eq 4 are comparable with Becher's values, as summarized in Table 2.

$$\log_{10}\text{cmc} = (1.646 \pm 0.082) - (0.496 \pm 0.008)\text{C\#} + (0.0437 \pm 0.0094)\text{EO\#} \quad (4)$$

Given the successful results presented by Ravey,[4] which showed an improved correlation by adding the product of C# and EO# to eq 4, some additional nonlinear terms were tested. We found that among our two-parameter models, the best fit is not the simple pair C# and EO#, but the combination of the carbon number C# and the cross term C#EO# ($R^2 = 0.996$, $F = 3145$, $s^2 = 0.0123$). This two-parameter model (eq 5) is statistically comparable to Ravey's three-parameter model ($R^2 = 0.996$, $F = 2209$, $s^2 = 0.0117$).

$$\log_{10}\text{cmc} = (1.902 \pm 0.075) - (0.523 \pm 0.009)\text{C\#} + (0.00441 \pm 0.00071)\text{C\#EO\#} \quad (5)$$

Finally, the CODESSA heuristic method has been applied to the same set of 31 linear alkyl ethoxylates, using the set of 400+ CODESSA-calculated descriptors only (excluding C#- and EO#-based descriptors). This analysis did not offer a better correlation model for the linear alkyl ethoxylates, even when CODESSA descriptors were combined with the descriptors based on C#, EO#, and related nonlinear combinations. This is not unexpected as the statistical fitness of both eqs 4 and 5 covers the experimental imprecision, and thus the additional descriptors would reflect only the experimental errors.

Prediction of cmc for the Diverse Set of Nonionic Surfactants. The CODESSA heuristic method of selection of the best regression models has been applied to the complete set of 77 nonionic surfactants (see Tables 8 and 9). The preliminary regression analysis was done using all original CODESSA descriptors, with the descriptors calculated for the entire molecule. This analysis revealed that the logarithm of the cmc is primarily determined by the hydrophobic part of the surfactant, and the quantum chemical and electrostatic descriptors do not perform better for this property than do simple constitutional and topological descriptors. When descriptors for the hydrophobic domains were generated, using the fragment descriptor calculation feature in the CODESSA program, the regression results were greatly improved. The hydrophobic fragments were defined as shown with dashed lines in Figure 1.

The plot of $\log_{10}$ cmc vs the calculated fractional H-bonding surface area of the molecules (Figure 2) reveals the clustering of the different nonionic surfactant classes. All ethoxylate classes fall within the same cluster, as do the amides, while the mono- and disaccharides form separate clusters. The H-bonding surface area includes contributions from nitrogen and oxygen atoms, as well as the hydrogen atoms bound to them. For a given class, cmc values decrease as the number of carbon atoms in the hydrophobic fragment increases.

The best correlation model (Table 4) includes two topological descriptors for the hydrophobic fragment (prefix "c-") and one descriptor for the hydrophilic contribution. The scatter plot (calculated vs experimental $\log_{10}$ cmc) for this regression is presented in Figure 3. The cmc can be effectively predicted for a diverse set of nonionic surfactants using eq 6.

$$\log_{10}\text{cmc} = -(1.80 \pm 0.16) - (0.567 \pm 0.009)\text{c-KH0} + (1.054 \pm 0.048)\text{c-AIC2} + (7.5 \pm 1.0)\text{RNNO} \quad (6)$$

The Kier & Hall index of 0 order for the hydrophobic fragment (c-KH0) represents the size of the hydrophobic fragment and contains group contributions from all nonhydrogen atoms in the fragment. This descriptor value increases with the size of the fragment, which causes a decrease in cmc, due to the negative descriptor coefficient.

(39) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*, 2nd ed.; Irwin: Homewood, IL, 1985.
(40) Puvvada, S.; Blankschtein, D. *J. Chem. Phys.* **1990**, *92*, 3710.
(41) Nagarajan, R.; Ruckenstein, E. *Langmuir* **1991**, *7*, 2934.

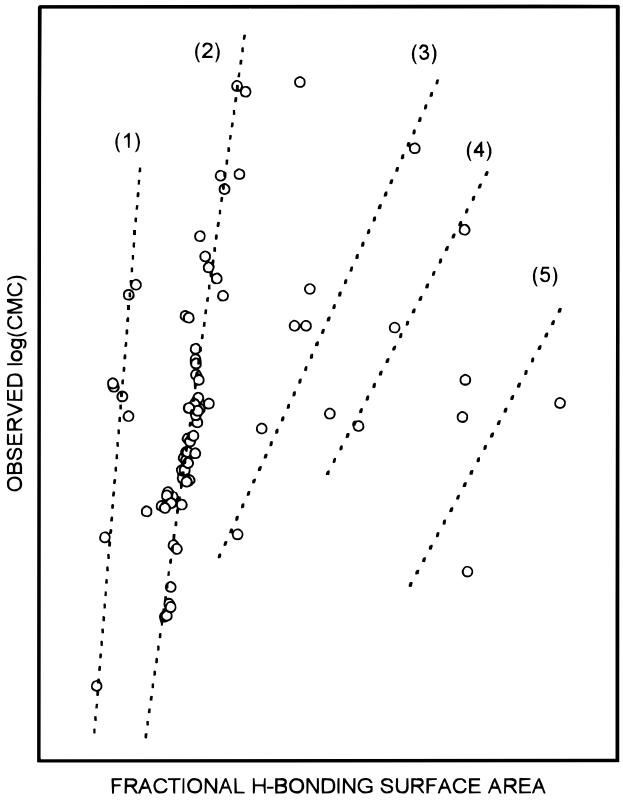

Figure 2. Observed $\log_{10}$ cmc vs the fractional hydrogen-bonding surface area, showing clustering of nonionic surfactant classes. The classes in each cluster are (1) ethoxylated alkylamides and fluorinated ethoxylated amides, (2) linear, branched, octylphenyl, fluorinated and alkylamine ethoxylates, (3) alkanediols, (4) alkyl monosaccharides, and (5) alkyl disaccharides. In each class, log cmc is decreasing with increasing carbon number.

Table 4. Details of the Best Correlation Model (R^2 = 0.9833, F = 1433, s^2 = 0.0313)

i	c_i	dc_i	t-test	descriptor
0	−1.802	0.1617	−11.14	intercept
1	−0.567	0.0093	−60.92	c-KH0
2	1.054	0.0477	22.09	c-AIC2
3	7.511	1.0124	7.42	RNNO

The average information content of order 2 for the hydrophobic fragment (c-AIC2) represents the complexity of the hydrophobic fragment and increases with the structural diversity of the fragment, which in turn causes an increase in cmc. The final descriptor, the relative number of nitrogen and oxygen atoms (RNNO), represents the size of the hydrophilic fragment and accounts for the increase in cmc with increasing EO number. The addition of a fourth descriptor to the model was not statistically justified, resulting in a very minor improvement of the correlation coefficient and standard error and a decrease of the F-test statistic, as can be seen in Table 5.

Calculation of Descriptors. The Kier & Hall index of 0 order (KH0) is defined by the following formula[7]

$$\mathrm{KH0} = \sum_{i=1}^{N} (\delta_i^v)^{-1/2} \quad \text{where} \quad \delta_i^v = \frac{Z_i^v - H_i}{Z_i - Z_i^v - 1} \qquad (7)$$

where Z_i is the total number of electrons in the ith atom, Z_i^v is the number of valence electrons, and H_i is the number of hydrogens directly attached to the ith atom. Valence contributions are summed for all atoms in the molecule or fragment, with the exception of the hydrogen atoms ($N = N_{TOT} - N_H$). Group contributions to KH0 are summarized in Table 6. KH0 values for the hydrophobic fragments (c-KH0) are summarized in Table 7.

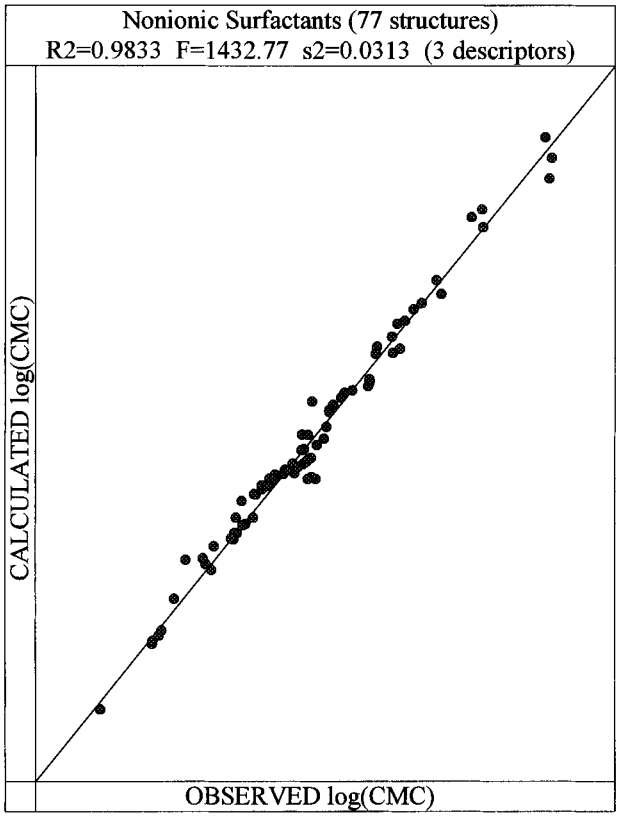

Figure 3. Scatter plot of the calculated vs observed $\log_{10}$ cmc for the best three-parameter model. The three descriptors are the Kier & Hall index of 0th order, the average information content of 2nd order, and the relative number of nitrogen and oxygen atoms.

Table 5. Best 77 Molecule Correlation Models and Their Statistical Characteristics

model	R^2	R^2_{cv}	F	s^2
best one-parameter model	0.8941	0.8889	633	0.1935
best two-parameter model	0.9689	0.9666	1151	0.0576
best three-parameter model	0.9833	0.9812	1433	0.0313
best four-parameter model	0.9849	0.9823	1172	0.0288

Table 6. Valence δ Values for Calculation of Kier & Hall Indices

group	$(\delta_i^v)^{-1/2}$
C atom from CH_2 group	0.707
C atom from CF_2 group	0.500
F atom from CF_2, CF_3 group	0.378
C^1 atom from phenyl group	0.500
C^2 atom from phenyl (CH)	0.577
CH_2 group	0.707
CH_3 group	1.000
C "group"	0.500 = 0.71 × CH_2
CH group	0.577 = 0.82 × CH_2
S atom	1.225 = 1.73 × CH_2
CF_2 group	1.256 = 1.77 × CH_2
phenyl group	3.308 = 4.68 × CH_2
CF_3 group	1.634 = 1.63 × CH_3

The information content indices are defined on the basis of the Shannon information theory and calculated as follows[9]

$$\mathrm{AIC2} = -\sum_{i=1}^{N_{class}} \frac{n_i}{n} \log_2 \frac{n_i}{n} \qquad (8)$$

Table 7. Hydrophobic Fragment Contribution to cmc and Estimated Alkane Carbon Number[a]

fragment label	structure	c-KH0	c-AIC2	HFOB	C#	C#$_{est}$
		Linear Alkanes				
C4	$C_4H_9-[O-]$	3.121	2.565	0.935	4	4.0
C6	$C_6H_{13}-$	4.536	2.524	0.090	6	5.8
C8	$C_8H_{17}-$	5.950	2.333	−0.912	8	8.0
C9	$C_9H_{19}-$	6.657	2.248	−1.403	9	9.0
C10	$C_{10}H_{21}-$	7.364	2.172	−1.884	10	10.1
C11	$C_{11}H_{23}-$	8.071	2.103	−2.356	11	11.1
C12	$C_{12}H_{25}-$	8.778	2.042	−2.821	12	12.1
C13	$C_{13}H_{27}-$	9.485	1.987	−3.280	13	13.0
C14	$C_{14}H_{29}-$	10.192	1.938	−3.733	14	14.0
C15	$C_{15}H_{31}-$	10.900	1.893	−4.181	15	15.0
C16	$C_{16}H_{33}-$	11.607	1.852	−4.625	16	15.9
		Linear Alkanediols				
C8D	$C_8H_{17}-[CH(OH)-]$	5.950	1.857	−1.414		9.1
C10D	$C_{10}H_{21}-[CH(OH)-]$	7.364	1.746	−2.333		11.0
C12D	$C_{12}H_{23}-[CH(OH)-]$	8.778	1.658	−3.226		12.9
		Branched Alkanes				
IC4	$(CH_3)_2CHCH_2-[O-]$	3.285	2.200	0.457		5.1
IC6	$(C_2H_5)_2CHCH_2-$	4.699	2.695	0.178		5.7
IC8	$(C_3H_7)_2CHCH_2-$	6.113	2.744	−0.572		7.3
IC10	$(C_4H_9)_2CHCH_2-$	7.527	2.744	−1.373		9.0
		Fluorinated Linear Alkanes				
CF6C	$C_6F_{13}CH_2-[C(O)N-]$	8.621	2.426	−2.327		11.0
CF6ESE	$C_6F_{13}C_2H_4SC_2H_4-[O-]$	11.967	3.290	−3.313		13.1
CF8C	$C_8F_{17}CH_2-[C(O)N-]$	11.133	2.248	−3.939		14.4
CF10C	$C_{10}F_{21}CH_2-[C(O)N-]$	13.644	2.103	−5.514		17.8
		Other Hydrophobes (Octylphenyl, Oleate)				
C8PHE	$t-C_8H_{17}-C_6H_4-[O-]$	10.017	2.862	−2.659		11.7
C18:1	$C_8H_{17}CH{=}CHC_7H_{14}-[C(O)O-]$	12.054	2.668	−4.018		14.6

[a] Note that atoms in brackets are not included in the fragment.

where n_i is the number of atoms in the ith class, n is the total number of atoms in the molecule or fragment, and N_{class} is the number of classes. The division of atoms into different classes depends upon the coordination sphere (order) taken into account. For a first-order information content calculation, atoms fall into the same class if (1) they are of the same type and valence, (2) they have the same number of first neighbors, (3) their first neighbors correspond in type and valence, and (4) they are connected to their corresponding first neighbors by bonds of the same multiplicity. For molecular fragments, this definition will result in atoms outside of the fragment influencing the determination of the number of classes. AIC2 values for the hydrophobic fragments (c-AIC2) are summarized in Table 7.

For example, let us consider calculation of the c-AIC2 of the simplest surfactant in the database, C_4E_1. There are a total of 13 atoms in this fragment. There are seven classes of atoms present, four classes of carbon atoms and three classes of hydrogen atoms. The classes are determined on the basis of the number and type of second neighbors of each atom. The four C classes each have one atom ($n_i = 1$). The number of second neighbors and type of neighbor, starting at the α carbon are the following: 4, C_2H_2; 6, COH_4; 6, CH_5; 3, CH_2. For the three H classes, all H atoms have three second neighbors. Two H atoms have COH, three have CH_2, and four have C_2H neighbors. Summing the contributions in eq 8 for $n = 13$ and $n_i = 1$, 1, 1, 1, 2, 3, 4 results in a c-AIC2 value of 2.565.

The final descriptor, RNNO, is calculated by simply dividing the number of these atoms by the total number of atoms in the molecule.

Application of the Best Correlation Model to the Linear Alkyl Ethoxylates. We have applied the best three-parameter model (eq 6) calculated using the entire 77 structure set, to the subset of 31 linear alkyl ethoxylates. The model demonstrated an excellent fit ($R^2 = 0.996$, $F = 2349$, $s^2 = 0.0110$), even though it was derived from the diverse 77 molecule data set. This compares very favorably with the model (eq 5) derived only from the 31 linear alkyl ethoxylates ($R^2 = 0.997$, $F = 2974$, $s^2 = 0.0087$).

Interpretation of Descriptors. The successful correlation achieved by using topological descriptors calculated for the hydrophobic fragment, which captures the nature of the hydrophobic regions of the surfactant molecule separately from the hydrophilic regions, is well supported by theoretical considerations of micellization. It is well-known that the formation of a micelle leads to the creation of a hydrophobic microenvironment in the core of the micelle. The loss of entropy that occurs by the ordering of the surfactant molecules in a micelle is less significant than the gain in configurational entropy of the water molecules caused by the removal of the hydrophobic domains from solution during the process of micellization.[5] In other words, the micellization potentials of amphiphiles depend essentially on the water volume displaced by the hydrophobic fragment.[4]

For linear molecules, the increase of the volume of the hydrophobic chain is simply related to the alkane carbon number. For other surfactants, branching of the hydrophobic fragment and presence of heteroatoms should be taken into account. The KH0 is a sum of valence δ's (δ_i^v) assigned to all skeleton atoms (i.e., all atoms excluding hydrogens) in a hydrophobic fragment. These valence δ's account for skeleton branching and electronic differences of heteroatoms (see eq 7). As an example, we can calculate that the CF_2 group is equivalent to 1.77 × CH_2 (Table 6). The c-KH0 descriptor increases with increasing number of non-hydrogen atoms in the hydrophobic fragment. The regression coefficient for this descriptor is negative, as expected, predicting a lower cmc for increasing c-KH0 values. The relative importance of the three descriptors can be seen from their individual correlation coefficients with log cmc. The r^2 values are 0.856, 0.123, and 0.009 for c-KH0, c-AIC2, and RNNO, respectively.

Table 8. Names of 77 Nonionic Surfactants by Class[a]

code	surfactant name	chemical formula
	Class: Ethoxylated Linear Alcohols (31 Structures)	
c4e#	butyl polyethylene oxide	$C_4H_9O(C_2H_4O)_mH$, $m = 1, 6$
c6e#	hexyl polyethylene oxide	$C_6H_{13}O(C_2H_4O)_mH$, $m = 3, 6$
c8e#	octyl polyethylene oxide	$C_8H_{17}O(C_2H_4O)_mH$, $m = 1, 3, 6, 9$
c10e#	decyl polyethylene oxide	$C_{10}H_{21}O(C_2H_4O)_mH$, $m = 3, 4, 6, 8, 9$
c11e8	undecyl octaethoxylate	$C_{11}H_{23}O(C_2H_4O)_8H$
c12e#	dodecyl polyethylene oxide	$C_{12}H_{25}O(C_2H_4O)_mH$, $m = 2-9, 12$
c13e8	tridecyl octaexthoxylate	$C_{13}H_{27}O(C_2H_4O)_8H$
c14e#	tetradecyl polyethylene oxide	$C_{14}H_{29}O(C_2H_4O)_mH$, $m = 6, 8$
c15e8	pentadecyl octaexthoxylate	$C_{15}H_{31}O(C_2H_4O)_8H$
c16e#	hexadecyl polyethylene oxide	$C_{16}H_{33}O(C_2H_4O)_mH$, $m = 6, 7, 9, 12$
	Class: Ethoxylated Branched Alcohols (5 Structures)	
ic4e6	propyl 2-methyl hexaethoxylate	$(CH_3)_2CHCH_2O(C_2H_4O)_6H$
ic6e6	butyl 2-ethyl hexaethoxylate	$(CH_3CH_2)_2CHCH_2O(C_2H_4O)_6H$
ic8e6	pentyl 2-propyl hexaethoxylate	$(CH_3CH_2CH_2)_2CHCH_2O(C_2H_4O)_6H$
ic10e#	hexyl 2-butyl polyethoxylate	$(CH_3CH_2CH_2CH_2)_2CHCH_2O(C_2H_4O)_mH$, $m = 6, 9$
	Class: Ethoxylated Octylphenols (10 Structures)	
c8phe#	*p,t*-octylphenyl polyethylene oxide	$C_8H_{17}(C_6H_4)O(C_2H_4O)_mH$, $m = 1-10$
	Class: Linear Alkane Diols (5 Structures)	
c8glycer	octyl α-glyceryl ether	$C_8H_{17}OCH_2CH(OH)CH_2OH$
c10diol	1,2-decanediol	$C_8H_{17}CH(OH)CH_2OH$
c11diol	1,3-undecanediol	$C_8H_{17}CH(OH)CH_2CH_2OH$
c12diol	1,2-dodecanediol	$C_{10}H_{21}CH(OH)CH_2OH$
c15diol	1,3-pentadecanediol	$C_{12}H_{25}CH(OH)CH_2CH_2OH$
	Class: Alkyl Glucose Ethers and Esters (7 Structures)	
c#gluc	*n*-alkyl-beta-D-glucoside, # = 8, 10, 12	$C_nH_{2n+1}O(C_6H_{11}O_5)$
c12delac	(*N*-dodecylamino)-1-deoxy-1-lactitol	$C_{12}H_{25}NH(C_6H_{12}O_4)O(C_6H_{11}O_5)$ (first ring open)
c12sucr	sucrose monolaurate	$C_{11}H_{23}C(O)O(C_6H_{10}O_4)O(C_6H_{11}O_5)$
c18sucr	sucrose monooleate	$C_8H_{17}CH{=}CHC_7H_{14}C(O)O(C_6H_{10}O_4)O(C_6H_{11}O_5)$
c12malt	beta-dodecyl maltoside	$C_{12}H_{25}O(C_6H_{10}O_4)O(C_6H_{11}O_5)$
	Class: Alkyl Amines and Amides (9 Structures)	
c11coneo	undecyl *N,N*-diethanolamide	$C_{11}H_{23}C(O)N(C_2H_4OH)_2$
c9cone#e	nonyl *N,N*-polyethoxyamide	$C_9H_{19}C(O)N[(C_2H_4O)_m)CH_3]_2$, $m = 3, 4$
c11cone#	undecyl *N,N*-polyethoxyamide	$C_{11}H_{23}C(O)N[(C_2H_4O)_m)CH_3]_2$, $m = 2, 3, 4$
c12Ze4	dodecyl Z-tetraethylene oxide *Z* = Ala (alanine, X = H, Y = CH_3) *Z* = Gly (glycine, X = H, Y = H) *Z* = Sar (sarcosine, X = CH_3, Y = H)	$C_{12}H_{23}N(X)CH(Y)C(O)O(C_2H_4O)_4H$
	Class: Fluorinated Linear Ethoxylates and Amides (10 Structures)	
cf6se#	(fluorinated linear ethoxylates)	$C_6F_{13}C_2H_4SC_2H_4(OC_2H_4)_mOH$, $m = 2, 3, 5, 7$
cf6se1se2	(fluorinated linear ethoxylates)	$C_6F_{13}C_2H_4SC_2H_4OC_2H_4SC_2H_4(OC_2H_4)_2OH$
cf6se#se#	(fluorinated linear ethoxylates)	$C_6F_{13}C_2H_4[SC_2H_4(OC_2H_4)_m]_2OH$, $m = 2, 3$
cf#cone3e3	(fluorinated *N*-ethoxylated amides)	$C_nF_{2n+1}CH_2C(O)N[(C_2H_4O)_3CH_3]_2$, $n = 6, 8, 10$

[a] Note that the ethoxylates have also been called polyethylene oxides, polyethylene glycols, and polyoxyethylenes in the literature.

Though the KH0 accounts for branching to some extent, it is still of limited applicability. To better take into account the branching and other structural features of hydrophobic fragments, the c-AIC2 is used. This descriptor takes into account the first and second neighbors of each atom and represents the relative structural diversity of a fragment compared to the maximum diversity possible for a structure with the same number of atoms. c-AIC2 decreases with an increase of the alkane chain of linear ethoxylates because the structural diversity of the hydrophobic fragment remains constant while the maximum possible diversity (i.e., the number of atoms) increases. However, c-AIC2 values for octylphenyl ethoxylates and thioethylene-containing surfactants are significantly higher than for the linear ethoxylates with the same number of atoms. The regression coefficient for c-AIC2 is positive, indicating that branching and other structural features serve to increase the cmc over that of a linear molecule with the same number of carbon atoms.

The dependence of the micellization on the hydrophilic part of the molecule is accounted for by the third parameter in the model, the RNNO. This descriptor accounts for the atoms in the hydrophilic fragment of the surfactant molecule that act as hydrogen bond acceptors. It has a positive coefficient value, indicating that the cmc increases with an increasing proportion of N and O atoms.

Fragment Selection. The CODESSA program can give insight into the proper selection of the hydrophilic and hydrophobic fragments. A good example of this came from the molecules containing sulfur. Ten structures from the database include at least one thioethylene group. Although it is noted that the sulfur in this group exhibits a slightly hydrophobic character,[20] we initially defined the SC_2H_4 group as a part of the hydrophilic fragment. The best correlation model had a squared correlation coefficient of $R^2 = 0.978$, using three descriptors. The addition of a fourth descriptor did not improve correlation significantly. The fourth descriptor, the number of sulfur atoms, was interesting because of its negative regression coefficient. This indicates that the SC_2H_4 group should be considered a part of the hydrophobic fragment. On recalculation of the fragment descriptors with the SC_2H_4 group in the hydrophobic fragment, the heuristic method was again applied to the entire set of 77 structures. A better fit was achieved ($R^2 = 0.983$), resulting in the final regression equation (eq 6).

Estimation of Alkane Carbon Number. The Kier & Hall connectivity index of the zeroth order and average information content of the second order together describe the essential influence of the hydrophobic fragment on micellization. The scatter plot (Figure 4) between the calculated and observed $\log_{10}$ cmc, using a calculation

Table 9. cmc Literature Values (25 °C) and Calculated Values for 77 Nonionic Surfactants[a]

structure codes	predicted $\log_{10}$ cmc	observed $\log_{10}$ cmc	Δ	ref
C4E1	−0.184	−0.009	−0.175	[26] Donbrow, 1966
C4E6	0.056	−0.110	0.166	[24] Elworthy, 1964
C6E3	−0.996	−1.000	0.004	[27] Corkill, 1964
C6E6	−0.877	−1.164	0.287	[24] Elworthy, 1964
C8E1	−2.272	−2.310	0.038	[28] Shinoda, 1959
C8E3	−2.088	−2.125	0.037	[27] Corkill, 1964
C8E6	−1.952	−2.004	0.052	[27] Corkill, 1964
C8E9	−1.880	−1.886	0.006	[27] Corkill, 1964
C10E3	−3.129	−3.222	0.093	[27] Corkill, 1964
C10E4	−3.070	−3.167	0.097	[30] Hudson, 1964
C10E6	−2.985	−3.046	0.061	[27] Corkill, 1964
C10E8	−2.926	−3.000	0.074	[2] Meguro, 1981
C10E9	−2.903	−2.886	−0.017	[27] Corkill, 1964
C11E8	−3.423	−3.523	0.100	[2] Meguro, 1981
C12E2	−4.198	−4.481	0.283	[31] Rosen, 1982
C12E3	−4.123	−4.284	0.162	[31] Rosen, 1982
C12E4	−4.063	−4.194	0.131	[31] Rosen, 1982
C12E5	−4.014	−4.194	0.180	[31] Rosen, 1982
C12E6	−3.974	−4.060	0.086	[27] Corkill, 1964
C12E7	−3.940	−4.086	0.146	[31] Rosen, 1982
C12E8	−3.912	−4.000	0.088	[31] Rosen, 1982
C12E9	−3.887	−4.000	0.113	[32] Lange, 1965
C12E12	−3.829	−3.854	0.025	[32] Lange, 1965
C13E8	−4.392	−4.569	0.177	[2] Meguro, 1981
C14E6	−4.930	−5.000	0.070	[27] Corkill, 1964
C14E8	−4.865	−5.046	0.181	[2] Meguro, 1981
C15E8	−5.333	−5.456	0.123	[2] Meguro, 1981
C16E6	−5.861	−5.780	−0.081	[33] Elworthy, 1962
C16E7	−5.825	−5.770	−0.055	[33] Elworthy, 1962
C16E9	−5.768	−5.678	−0.090	[33] Elworthy, 1962
C16E12	−5.703	−5.638	−0.065	[33] Elworthy, 1962
C8PHE1	−4.119	−4.305	0.186	[34] Crook, 1963
C8PHE2	−4.019	−4.116	0.097	[34] Crook, 1963
C8PHE3	−3.943	−4.013	0.070	[34] Crook, 1963
C8PHE4	−3.883	−3.886	0.003	[34] Crook, 1963
C8PHE5	−3.835	−3.824	−0.011	[34] Crook, 1963
C8PHE6	−3.795	−3.678	−0.117	[34] Crook, 1963
C8PHE7	−3.762	−3.602	−0.160	[34] Crook, 1963
C8PHE8	−3.734	−3.553	−0.181	[29] Crook, 1964
C8PHE9	−3.710	−3.523	−0.187	[29] Crook, 1964
C8PHE10	−3.689	−3.481	−0.208	[29] Crook, 1964
IC4E6	−0.422	−0.049	−0.373	[24] Elworthy, 1964
IC6E6	−0.789	−1.016	0.227	[24] Elworthy, 1964
IC8E6	−1.612	−1.670	0.058	[24] Elworthy, 1964
IC10E6	−2.474	−2.547	0.073	[24] Elworthy, 1964
IC10E9	−2.393	−2.526	0.133	[24] Elworthy, 1964
C8GLYCER	−2.121	−2.237	0.116	[28] Shinoda, 1959
C10DIOL	−2.774	−2.638	−0.136	[35] Kwan, 1980
C11DIOL	−2.810	−2.638	−0.172	[35] Kwan, 1980
C12DIOL	−3.759	−3.745	−0.014	[35] Kwan, 1980
C15DIOL	−4.721	−4.886	0.165	[35] Kwan, 1980
C8GLUC	−1.775	−1.602	−0.173	[36] Shinoda, 1961
C10GLUC	−2.851	−2.658	−0.193	[36] Shinoda, 1961
C12GLUC	−3.872	−3.721	−0.151	[36] Shinoda, 1961
C12DELAC	−3.151	−3.222	0.071	[23] Auvray, 1995
C12MALT	−3.603	−3.620	0.017	[23] Auvray, 1995
C12SUCR	−3.032	−3.469	0.438	[37] Herrington, 1986
C18SUCR	−4.881	−5.292	0.412	[37] Herrington, 1986
C11CONEO	−3.591	−3.585	−0.006	[5] Rosen, 1989
C9CONE3E	−2.463	−2.299	−0.164	[22] Selve, 1991
C9CONE4E	−2.414	−2.193	−0.221	[22] Selve, 1991
C11CONE2	−3.541	−3.398	−0.143	[22] Selve, 1991
C11CONE3	−3.468	−3.292	−0.176	[22] Selve, 1991
C11CONE4	−3.415	−3.611	0.197	[22] Selve, 1991
C12ALAE4	−3.940	−3.413	−0.527	[38] Kuwamura, 1984
C12GLYE4	−3.913	−3.474	−0.439	[38] Kuwamura, 1984
C12SARE4	−3.940	−3.533	−0.407	[38] Kuwamura, 1984
CF6CONE3	−3.328	−3.260	−0.068	[22] Selve, 1991
CF8CONE3	−4.999	−4.921	−0.077	[22] Selve, 1991
CF10CONE	−6.625	−6.523	−0.102	[22] Selve, 1991
CF6SE2	−4.646	−4.602	−0.043	[21] Matos, 1989
CF6SE3	−4.569	−4.553	−0.016	[21] Matos, 1989
CF6SE5	−4.462	−4.432	−0.030	[21] Matos, 1989
CF6SE7	−4.391	−4.319	−0.072	[21] Matos, 1989
CF6SESE2	−4.630	−4.638	0.008	[21] Matos, 1989
CF6SE2SE	−4.571	−4.585	0.014	[21] Matos, 1989
CF6SE3SE	−4.482	−4.469	−0.012	[21] Matos, 1989

[a] All concentrations are in units of moles/liter.

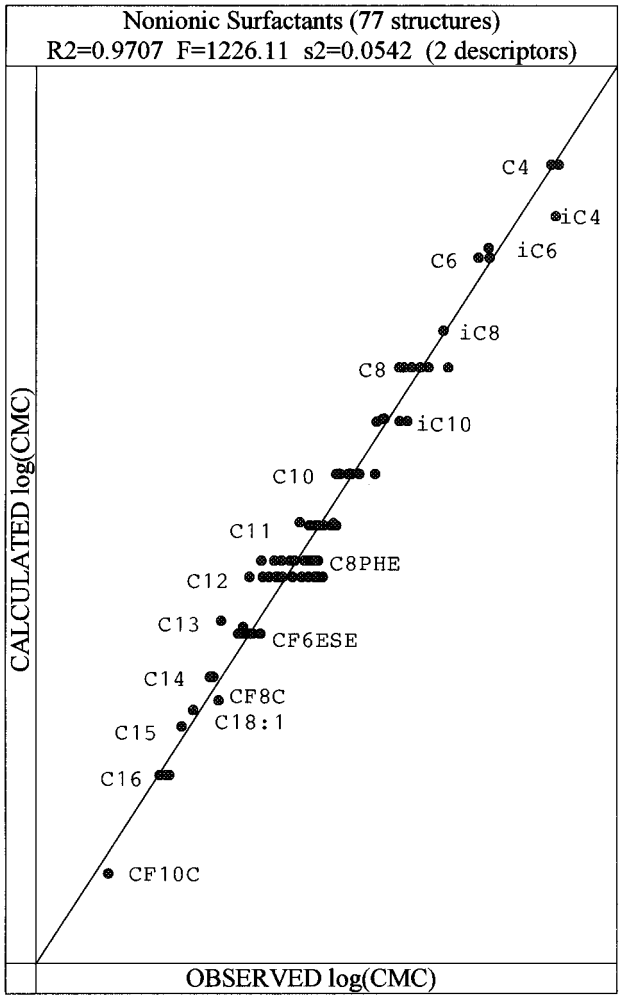

Figure 4. Scatter plot of the calculated vs observed $\log_{10}$ cmc for the two-parameter model involving only the hydrophobic fragment descriptors (Kier & Hall index of 0th order and average information content of 2nd order). See Table 7 for fragment label descriptions.

based only on the two hydrophobic descriptors, graphically shows that the calculated $\log_{10}$ cmc has a constant increment per each additional CH_2 group for linear ethoxylates. On the basis of this increment, it is possible to calculate CH_2 equivalents for other structures. Estimates of the alkane carbon number for any hydrophobic fragment can be calculated using the combined contributions from the two hydrophobic descriptors. The hydrophobic contribution to cmc (HFOB) is defined in eq 9 and tabulated for all fragments in this study (Table 7). Estimated alkane carbon numbers for any hydrophobic fragment can be calculated using eq 10.

$$\mathrm{HFOB} = -0.5666\text{c-KH0} + 1.054\text{c-AIC2} \quad (9)$$

$$\mathrm{C\#_{est}} = 6.037 - 2.134\mathrm{HFOB} \quad (10)$$

The regression for $\mathrm{C\#_{est}}$ was calculated ($r^2 = 0.9996$) using the HFOB values for the 11 linear alkane hydrophobic fragments, ranging in size from *n*-butyl to *n*-hexadecyl fragments. Estimated carbon numbers (also known as ACN, alkane carbon number, or EACN, equivalent alkane carbon number) are presented in Table 7. The results demonstrate that fragments with branching, double bonds, and aromatic ring structures behave as if they have fewer than the actual number of carbon atoms present (note the exception of the branched butyl fragment). The fluorinated alkanes, on the other hand, act

as if they have more carbons than the actual number. A quick estimate of the ACN for a fragment can be made by summing valence δ group contributions from Table 6. For best results, though, this estimate must be modified by the c-AIC2 contribution. As an example, consider the fluorinated $C_6F_{13}CH_2-$ fragment. By summing valence δ contributions, the ACN estimate is 11.5 but falls to 11.0 when the c-AIC2 contribution is considered. Similar results may be found for the octylphenyl fragment, with an ACN of 12.7 from the valence δ estimate and 11.7 when the c-AIC2 term is included. From this example it can be seen that the Kier & Hall index accounts for branching and ring structure by providing an ACN of less than the actual number of 14 carbon atoms for octylphenol, but for the best estimate, the additional c-AIC2 term is needed to better account for the structural complexity.

Simplified cmc Calculation. For any given nonionic surfactant, if the hydrophobic fragment of interest appears in Table 7, the topological descriptors do not need to be calculated to arrive at a cmc estimate. Given the precalculated HFOB values, the cmc can be estimated from eq 11, where N_{TOT} is the total number of atoms, N_N is the number of nitrogen atoms, and N_O is the number of oxygen atoms in the molecule.

$$\log_{10} \text{cmc} = \text{HFOB} + 7.51(N_N + N_O)/N_{TOT} \quad (11)$$

Conclusion

It is evident that the proposed three-parameter model (eq 6) can be used to predict the cmc for nonionic surfactants of diverse chemical structure with a significant degree of confidence. The model employs descriptors calculated only from the molecular topology, and cmc predictions can be made using a pocket calculator.

This QSPR study of the critical micelle concentration of nonionic surfactants employed the screening of a wide selection of molecular descriptors describing various molecular features, from constitutional to quantum chemical in nature. In agreement with previous studies, it was found that the cmc of nonionic surfactants in aqueous solution is primarily determined by the hydrophobic part of the molecule. The logarithm of the cmc decreases with an increase in the size of the hydrophobic fragment and increases with an increase in the relative size of the hydrophilic fragment. Hydrophobicity is affected by branching of the hydrophobic fragment and the presence of heteroatoms.

The success of this approach for the prediction of nonionic surfactant cmc values using CODESSA-generated fragment descriptors encourages us to continue this work for the prediction of cmc for other families of surfactants, namely the zwitterionic, anionic, and cationic surfactants, as well as the prediction of other surfactant properties.

Acknowledgment. The authors P.D.T.H. and D.O.S. thank Kraft General Foods, ICI Surfactants, and the National Science Foundation (Grant CTS 9215384) for their support of this research.

LA950581J