JOURNAL OF PHYSICAL ORGANIC CHEMISTRY
J. Phys. Org. Chem. 2004; **17:** 431–438
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/poc.749

Correlation and prediction of critical micelle concentration using polar surface area and LFER methods

Robert A. Saunders and James A. Platts*

Department of Chemistry, Cardiff University, P.O. Box 912, Cardiff CF10 3TB, UK

Received 11 October 2004; revised 8 January 2004; accepted 16 January 2004

epoc ABSTRACT: Models of critical micelle concentration (CMC) using two separate methods, the linear free energy relationship of Abraham and a modified polar surface area approach, are reported. Individual models are developed for anionic, non-ionic and structurally diverse molecules, the last including many commercially important drugs such as analgesics, anaesthetics and antibiotics. Statistical analysis demonstrates the predictive accuracy of both methods, with R^2 values around 0.90 throughout. A further model for the simultaneous calculation of CMC for anionic and non-ionic surfactants was developed, giving reasonable correlations of observed vs calculated CMC. Both methods show similar patterns in regression coefficients; the most significant factor affecting a molecule's CMC is its size, with larger surfactants giving lower CMC. Strong H-bond acidic surfactants form micelles at lower concentrations, and increasing the H-bond basicity of a surfactant acts to raise the CMC. Copyright © 2004 John Wiley & Sons, Ltd.
Additional material for this paper is available in Wiley Interscience

KEYWORDS: critical micelle concentration; polar surface area; LFER

INTRODUCTION

A molecule's critical micelle concentration (CMC) is defined as the concentration range at which individual isolated surfactant molecules begin to aggregate to form micelles due to surface activity. After the CMC is exceeded, any additional surfactant added to the solution will form micelles. Once the CMC of a surfactant has been reached, many important physicochemical properties such as surface tension, conductivity and detergency change dramatically.[1] These properties are important to many industrial and biological systems, so the ability to predict CMCs easily directly from molecules' structure is of great interest.

The relationship between molecular structure and CMC has been well documented. A typical surfactant molecule can be broken down to two components that contribute towards the CMC, namely the hydrophobic (tail) and hydrophilic region (head). As the size of the hydrophobic region is increased it becomes more thermodynamically favourable for the hydrophobic regions of the surfactant molecule to minimize contact with the aqueous solution, seen as a decrease in CMC. In contrast, as the size and hydrophilic properties of the head group are increased, the CMC rises.[1]

**Correspondence to:* J. A. Platts, Department of Chemistry, Cardiff University, P.O. Box 912, Cardiff CF10 3TB, UK.
E-mail: platts@cf.ac.uk
Contract/grant sponsor: UK Engineering and Physical Science Research Council; *Contract/grant number:* GR/N20638.

Linear relationships between logCMC (CMC typically measured in mol l^{-1}) and the number of carbon atoms in a surfactant's hydrophobic tail have been defined for homologous series of linear alkyl hexaethoxylates by Rosen[2] and octaethoxylates by Merguro *et al.*[3] Ravey *et al.*[4] showed a linear relationship between the number of ethylene oxide units and logCMC for dodecyl polythoxylates. Beecher[5] used both the number of carbon atoms and number of ethylene oxide units to predict CMC for a series linear alkyl ethoxylate surfactants. The following equation was produced:

$$\log\mathrm{CMC} = A + Bm + Cn \quad (1)$$

where m and n are the number of carbon atoms and ethylene oxide units, respectively, and A, B and C are regression coefficients. The predictive ability of this relationship was improved by Ravey *et al.*,[4] who introduced a non-linear descriptor, a cross term defined as the number of carbon atoms multiplied by number of ethylene oxide units.

There has also been great deal of success in predicting CMC using quantitative structure–property relationships (QSPR). Wang *et al.*[6] derived the following equation for a set of 29 linear alkyl ethoxylates and 10 alkyl phenyl polyethylene oxides:

$$\log\mathrm{CMC} = 1.930 - 0.7846\mathrm{KH0} - 8.871 \times 10^{-5}E_{\mathrm{T}} + 0.04938D \quad (2)$$
$$N = 39, R^2 = 0.995$$

where KH0 is the Kier and Hall index of zeroth order,[7] E_T is the total molecule energy (in eV) and D is the dipole moment (in debye) of the surfactant and N is the number of molecules that were used in the regression. Direct comparison of Eqn (2) and those proposed by Ravey *et al.*[4] and Beecher[5] revealed that Eqn (2) was as accurate as the previous models but had the benefit that it could be used to predict CMC not only for alkyl ethoxylates but also alkyl phenyl polyethylene oxides.

Huibers *et al.*[8] used the program CODESSA[9] (Comprehensive Descriptors for Structural and Statistical Analysis) to predict CMC for a series of 77 non-ionic surfactants. The CODESSA program uses a heuristic approach to select the most appropriate descriptors from a large pool of several hundred descriptors. The study produced the following equation:

$$\begin{aligned}\text{logCMC} &= -1.802 - 0.567\text{c-KH0} + 1.054\text{c-AIC2} \\ &\quad + 0.751\text{RNNO} \qquad (3)\\ N &= 77, R^2 = 0.983\end{aligned}$$

where AIC2 is the information content index,[10] RNNO (relative number of nitrogen and oxygen) is the number of oxygen and nitrogen atoms divided by the total number of atoms in the molecule and the prefix c- indicates that the descriptor refers only to the hydrophobic regions of the surfactant.

Huibers *et al.* followed up this study by using CODESSA to derive an equation for the prediction of CMC for anionic surfactants.[11] This equation was based on a dataset of 119 sulfonates and sulfate molecules. CODESSA produced the following equation:

$$\begin{aligned}\text{logCMC} &= 1.89 - 0.314\text{t-sum-KH0} \\ &\quad - 0.034\text{TDIP} - 1.45\text{h-sum-RNC} \qquad (4)\\ N &= 119, R^2 = 0.940\end{aligned}$$

where t-sum-KH0 is the Kier and Hall molecular connectivity indices of zeroth order[7] for all hydrophobic regions, TDIP is the total dipole of the molecule, and h-sum-RNC is the sum of the relative number of carbon atoms for hydrophilic regions.

The R^2 value for Eqns (2)–(4) show that the predictive accuracy of these models is of a high quality. However, these models are all constructed from datasets with low diversity of functional groups; for instance, many of the molecules within these datasets are homologous series. The aim of this study was to try to establish a more general model for the prediction of CMC for more structurally diverse molecules such as drug molecules.

While the heuristic approach of programs such as CODESSA may find correlations that could otherwise have been missed, the models produced often forgo the clarity and interpretability of models produced using other QSPR methods. A further aim for this study is that from the models produced further information can be inferred about the physiochemical factors influencing the formation of micelles.

We chose to create two separate models using two different QSPR methods. The first of these is the LFER of Abraham *et al.*,[12,13] which splits all important solute–solvent interactions into five physiochemical descriptors. These descriptors are defined as follows:

E = the molar refraction of the solute minus the molar refraction of an alkane of equivalent volume
S = the combined dipolarity/polarizability of the molecule
A = the total hydrogen bonding acidity for the molecule
B = the total hydrogen bonding basicity for the molecule
V = McGowan's characteristic molecular volume.[14]

These descriptors are combined to form the following linear equation:

$$\text{logSP} = c + eE + sS + aA + bB + vV \qquad (5)$$

where logSP is the logarithm of a solvation property and c, e, s, a, b and v are regression coefficients and can be regarded as constants for a given system. It is these coefficients that contain the complementary effects of the phase on the interactions for a given system. The relative size of these coefficients represents the importance and function of its associated descriptor within the system.

Descriptors are calculated using the group contribution method of Platts *et al.*[15] In this method, a molecule is broken down into its component fragments, each of which has an associated value for each descriptor; the values of these fragments are then summed to give the descriptor value for the molecule. This method has been successfully applied to a wide variety of chemical, biological and environmental systems such as water–octanol partition,[16] solubility in supercritical CO_2[17] and blood–brain distribution.[18]

The second method that was chosen to predict CMC is the surface area approach of Saunders and Platts.[19] The method uses a set of descriptors derived from PSA (polar surface area). Traditionally, PSA is defined as the van der Waals surface area of all nitrogen and oxygen atoms and hydrogen attached to nitrogen or oxygen atoms.[20] PSA has been of a great interest since its introduction in 1990[21] and has been used successfully to model many important properties such as blood–brain barrier partition,[22] intestinal absorption[23] and oral bioavailability.[24] However, despite these successes, PSA is not without problems.

The first problem in using PSA as a descriptor is that it reduces the various ways in which a molecule can interact with its environment to a single number. Work by Abraham[12] and others has demonstrated that the relative

importance of these interactions can differ greatly. Stenburg *et al.*[25] proposed a deconvoluted version of PSA, denoted PTSA, which was used to predict intestinal absorption and shown to be superior to that of traditional PSA equations. The PTSA definition split PSA into several individual atom-type surface area descriptors that included sp^3 hybridized nitrogen, double-bonded oxygen and singly bonded oxygen.

In the method of Saunders and Platts,[19] this problem has been addressed by splitting PSA into its component H-bond acid (ASA) and H-bond base surface area (BSA). ASA is defined as the surface area of all hydrogen atoms attached to nitrogen or oxygen; BSA is defined as the surface area of all nitrogen and oxygen atoms. Other surface area descriptors based on molecular surface areas used within this method are total molecular surface area (TSA), total aromatic carbon surface area (BenSA) and total halogen surface area (HalSA); the last two surface areas are included to account for the polar and polarizability properties of a molecule. Figure 1 demonstrates how the surface area of a surfactant is partitioned by these descriptors. By analogy with Eqn (5), the following equation was proposed:

$$\begin{aligned}\log SP = c + t\mathrm{TSA} + a\mathrm{ASA} + b\mathrm{BSA} \\ + h\mathrm{HalSA} + e\mathrm{BenSA}\end{aligned} \quad (6)$$

The coefficients c, t, a, b, h and e are akin to those in Eqn (5), being derived from regression and their values representing the physical properties of the system.

The second problem of using PSA as a descriptor is that it does not account for the varying H-bonding properties of different functional groups. One solution to this was the method of Winiwarter *et al.*,[26] in which molecular surface areas were scaled using partial atomic charge. This problem has been tackled by Saunders and Platts[19] by defining 46 simple molecular fragments, each assigned a scaling factor based on their position in the Abraham A or B[12] scales. The inclusion of these scaling factors meant that the definition of PSA could be expanded to include sulfur and phosphorus atoms. These new surface area descriptors have been shown to yield more flexible and accurate predictive models for octanol–water partition ($\log P_{oct}$) chloroform–water partition ($\log P_{chcl}$) and cyclohexane–water partition ($\log P_{cyc}$).[19]

Although the method of Abraham and the modified polar surface area approach were developed for transfer processes involving one or more liquid, solid, or solution phases, we hypothesize that these methods will be flexible enough to model CMC. An analogy can be made between CMC and two-phase partition processes, except that the partition is between solvated and associated solute. CMC is determined by factors such as the self-association properties of water and the relative hydrophobicity of the tail, properties that can be accounted for using descriptors such as molecular size and volume. CMC is also determined by the self-association and repulsive properties of the surfactant head groups, which can be accounted for with descriptors that encompass the molecules hydrogen bonding abilities.

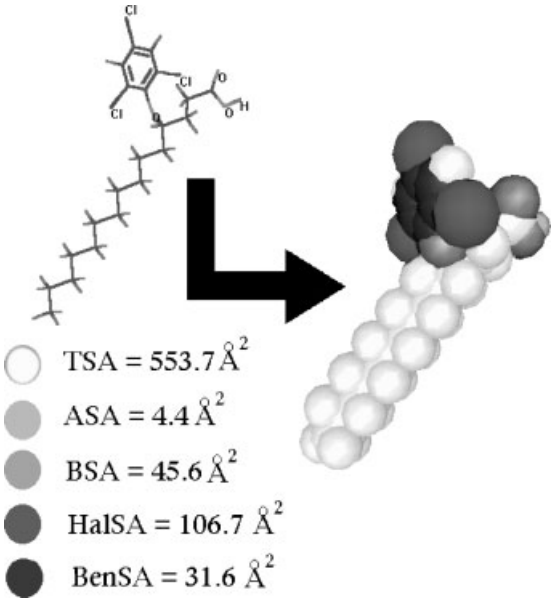

Figure 1. Partitioning of the surfactant surface area

METHOD

Three separate data sets were compiled, the first two from previous studies of CMC by Huibers *et al.*[8,11] Dataset 1 contained 77 non-ionic surfactants in aqueous solution at 25 °C with logCMC values ranging from −6.523 to −0.009 log units. Dataset 2 contained 119 anionic surfactants in aqueous solution at 40 °C; 50 of these values were recorded at 25 °C and their values at 40 °C were calculated using the recommended ratios of 1.088 and 1.030 for sulfonates and sulfates, respectively, which have been established to be approximately constant for the CMC of these molecules.[11] For dataset 2, logCMC values ranged from −4.899 to −0.496 log units.

A third data set was compiled from Schreier *et al.*'s paper.[27] This dataset contains 32 drug molecules in aqueous solution at 30 °C. These molecules include analgesics, anaesthetics and antibiotics, the logCMC values of which range from −6.22 to −0.60 log units. It should be noted that all of these molecules have been seen to form micelles and do not associate in a manner in which aggregate size increases continuously with increasing concentration. A comprehensive list of all the molecules from all datasets with their associated logCMC values has been deposited as Supplementary Data, available in Wiley Interscience.

The first two datasets allow direct comparison between the methods used in this study and previously published ones. Also, dataset 1 and the majority of dataset 2 can be

combined, allowing our methods to be applied simultaneously to the calculation of CMC for ionic and non-ionic surfactants. Dataset 3 was selected as it contains many drug molecules that are already of great commercial interest, and also the number of different functional groups present is significantly broader than that of any previous study of the prediction of CMC.

All of the molecules in all datasets were converted into SMILES[28] (Simplified Molecular Input Line Entry Specification). Abraham descriptors were then calculated from these SMILES using the group contribution method of Platts *et al.*[15] running on an SGI O^2 computer. It should be noted that the group contribution of Platts *et al*. does not contain fragments for the anionic oxygen of the sulfonates and sulfate in dataset 2, so the SMILES were changed so the anionic oxygen was treated as an oxygen in S═O (this can be justified as every molecule in the dataset contains one anionic oxygen and hence can be regarded as constant within regression analysis).

Surface area descriptors were calculated by generating initial approximate 3D descriptors from SMILES strings using the program CORINA.[29] These approximate 3D structures were then energy minimized using MM+ running in HYPERCHEM 6[30] with optimization terminating when < 0.01 kcal. An in-house modified version of the program MOLVOL[31] was then used to calculate surface area descriptors. This modified version of MOLVOL includes the ability to read MDL mol files; from the connectivity data contained within the mol file the weighting factors for H-bonding acidity and basicity are allocated. Again, it should be noted that there is no scaling factor for the anionic oxygen present in dataset 2; solutions to this problem are discussed later.

Using these descriptors, coefficients were calculated via multiple linear regression analysis (MLRA) for each of the three datasets for both methods. The MLRA was carried out using the JMP statistical package published by SAS Software.[32] The accuracy and precision of the models were evaluated using the following statistical methods. Root mean square error (RMSE) was used to test the accuracy of the CMC values predicted by our models. *T*-ratios were used to determine the significance and importance of each of the descriptors within the regression. R^2 and cross-validated R^2 (R_{cv}^2) were used to evaluate the internal self-consistency of the models. R^2 is a statistical indication of the percentage of variance in the original dataset that is being modelled successfully.

RESULTS

Dataset 1: non-ionic

Regression of the CMC values of dataset 1 against scaled surface area descriptors gave the following equation:

$$\begin{aligned} \log\mathrm{CMC} &= 1.282 - 0.017\mathrm{TSA} - 0.256\mathrm{ASA} \\ &\quad + 0.067\mathrm{BSA} - 0.007\mathrm{HalSA} - 0.001\mathrm{BenSA} \\ N &= 77, R^2 = 0.903, \mathrm{RMSE} = 0.434, \\ R_{cv}^2 &= 0.880, F = 131.5 \end{aligned} \tag{7}$$

When the same regression is performed using the traditional PSA descriptor along with TSA, a significantly poorer model is produced with R^2 dropping by 0.65 and the RMSE rising by 0.736 log units. Table 1 contains the results of statistical analysis for all models. The drop in predictive accuracy is easily clarified when the *t*-ratios of the descriptors in Eqn (7) are analysed. The *t*-ratios show that ASA and BSA have equal but opposing effects on CMC, i.e. as ASA is increased the value of logCMC predicted by Eqn (7) will decrease whereas raising BSA serves to increase values of logCMC calculated by Eqn (7). Hence the amalgamation of ASA and BSA to form PSA will create a descriptor that cannot correctly account for the physiochemical properties of CMC, as determined by Eqn (7).

Table 1. Statistical analysis of models

Model	Descriptors	*n*	R^2	RMSE	R_{cv}^2	*F*-ratio
Dataset 1: non-ionic	TSA, PSA	77	0.25	1.179	0.174	12.2
	TSA, ASA, BSA, HalSA, BenSA		0.903	0.433	0.879	131.6
	Abraham LFER		0.856	0.527	0.805	84.6
Datset 2: anionic	TSA, PSA	119	0.851	0.338	0.839	335.3
	TSA, ASA, BSA,[b] HalSA, BenSA		0.871	0.320	0.855	151.9
	Abraham LFER		0.868	0.324	0.846	148.0
Combined data	TSA, PSA	127	0.39	0.998	0.350	39.9
from datasets	Scaled TSA, ASA, BSA,[a] HalSA, BenSA		0.757	0.757	0.741	75.3
1 and 2	Scaled TSA, ASA, BSA,[b] HalSA, BenSA, O-SA		0.826	0.543	0.810	114.8
	Scaled TSA, ASA, BSA,[b] HalSA, BenSA, indicator		0.815	0.570	0.800	94.9
Dataset 3: drug molecules	TSA, PSA	32	0.734	0.691	0.422	66.5
	Scaled TSA, ASA, HalSA, BenSA		0.909	0.418	0.829	67.7
	Abraham LFER		0.909	0.420	0.080	67.2

[a] O^- surface area included with value scaled to one.
[b] O^- surface area not included.

A regression of the same dataset against the Abraham descriptors produced the following equation:

$$\begin{aligned} \log CMC &= 1.2066 - 1.400E + 4.06S - 3.480A \\ &\quad + 1.522B - 3.148V \\ N &= 77, R\hat{}2 = 0.856, RMSE = 0.527, \\ R_{cv}^2 &= 0.805, F = 84.6 \end{aligned} \tag{8}$$

The statistics of this regression are similar to, but slightly worse than, those of Eqn (7) with a slight decrease in R^2 and a slight increase in RMSE. Although the statistics of Eqns (7) and (8) show both methods produce accurate models, the predictive power of both equations is less than that published by Huibers *et al.*[8] for the same dataset, Eqn (3). The main source of this loss of accuracy is that many molecules in the dataset contain large quantities of intramolecular hydrogen bonding, which has been seen to 'tie up' both acid and base atoms and thus alter their acid and base properties. Although the scaled surface area method contains simple definitions to account for intramolecular H-bonding around aromatic rings, the definitions are not comprehensive enough to calculate accurately the properties of molecules such as sucrose monooleate and β-dodecyl maltoside, which are seen to be the two largest outliers for the scaled surface method (residuals of -1.037 and 0.967, respectively). Work is ongoing to identify important classes of intramolecular H-bonding and their effects on *A*, *B* and PSA.[33]

Dataset 2: ionic surfactants

Table 1 contains various statistical analyses of the models produced from the 119 ionic surfactants of dataset 2. As there are no defined experimental values for the anionic oxygen of the sulfonates and sulfates in the Abraham scales of *A* and *B*, the following measures were taken to account for this in the scaled surface area method:

1. Models were created where the O^- surface area was incorporated into the definition of BSA and scaled with a value of one. This made it approximately equivalent to oxygen in sulfoxide.
2. The O^- surface area was removed from the definition of BSA and allocated its own descriptor, which was termed O^-SA.
3. The O^- surface area was completely omitted from the BSA descriptor.

The results showed that the above three methods make very little difference to the statistics of the model, with R^2 values of 0.866, 0.875 and 0.871 for methods 1, 2 and 3, respectively. This is because the surface areas of O^- are almost constant through the data set with values only ranging from 16.2 to 20.9 Å^2 with a standard deviation of 1.79 Å^2. Not only are the surface areas of the O^- constant but also their occurrence, with one present for each surfactant in the dataset.

In this case, the models made containing only PSA and TSA are only marginally worse than those made with the five-parameter scaled surface descriptors. This major change is due to the fact that 97% of molecules in dataset 1 have some H-bond acidity, whereas in dataset 2 only 16% of the surfactants contain H-bond acidic groups. Hence for dataset 2 PSA is dominated by BSA, and PSA and BSA are almost interchangeable. While the overall statistics of the models change very little, 14 of the 19 surfactants that do possess H-bond acidity show a marked improvement when ASA and BSA are used separately. Figure 2 shows the correlation between observed and calculated logCMC values for the five-parameter scaled surface area model.

The range of different functional groups that fall into the defined fragments of our scaling factors is very narrow, with only 12 different types occurring (two acid fragments and 10 base). The lack of structural diversity in the surfactants accounts for the fact that when the scaling factors for ASA and BSA are removed, the model produced is statistically comparable to that of the model that includes scaling factors.

The model produced by Huibers *et al.*[11] for this dataset gave an R^2 value of 0.94. The cause of the loss of accuracy here is the occurrence of numerous series of surfactants in which the tail group remains constant and the position of the head group moves from the terminal to the medial position along the carbon chain, e.g. 1-dodecanesulfate through to 6-dodecanesulfate. Using the LFER relationship approach of Abraham and the group contribtion method of Platts *et al.*, the descriptors for these surfactants will be calculated to be equal. Although the surface area method is 3D, the changes in descriptors for these series are so subtle that the associated changes in CMC are not modelled fully. Additional accuracy could be achieved by including topological descriptors such as KH0, but this would negate the physical interpretability of such a model and prevent

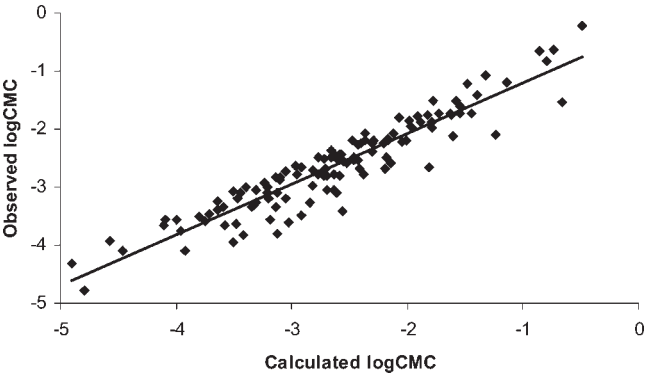

Figure 2. Observed logCMC values vs calculated logCMC for ionic dataset

comparison with other solvation phenomena, a key feature of methods based on the LFER method.

Although the results of these models show that molecular surface areas can predict CMC with reasonable accuracy, the nature of the dataset, with few surfactants displaying any hydrogen bond acidic properties and the highly similar structure of the molecules, does not challenge the PSA descriptor enough to merit its splitting into ASA and BSA or its scaling.

Combined dataset

One of the strengths of the molecular surface area models that we have produced is that it is possible easily to combine data from datasets 1 and 2 and simultaneously predict logCMC for anionic and non-ionic surfactants. The two datasets could not be combined directly as dataset 1 is measured at 25 °C and dataset 2 at 40 °C, and the temperature dependence of CMC is well documented.[34] Fifty surfactants from dataset 2 had CMC values recorded at 25 °C which could be combined with the 77 surfactants of dataset 1 which were also recorded at 25 °C. It is not possible to back-extrapolate the CMC of the remaining 69 surfactants in dataset 2 using the ratio stated earlier, as a number of the structures have Kraft points higher than 25 °C, meaning that micelles would not be formed at 25 °C and that calculated values would be physically meaningless. It not possible to combine the surfactants of dataset 3 as their CMC values were observed at 30 °C, and no single ratio can be assigned to such a diverse set.

MLRA was performed against this combined dataset of 127 CMC and their molecular surface area descriptors. The statistical analysis for these models is shown in Table 1. The models that employ only the PSA and TSA descriptors are clearly incapable of modelling CMC. Small improvements result from separating PSA, but acceptable statistics only result when the ASA and BSA are scaled to account for their H-bonding strengths, yielding an increase of ca. 25% in R^2 and a drop of 0.13 in RMSE over unscaled models.

If the anionic oxygen surface area is removed from BSA and included as a separated descriptor O^-SA, further improvement is seen in the model, causing R^2 and $R_{cv}{}^2$ to increase by 7% and RMSE to decrease by 0.21. Within this combined dataset, the O^-SA descriptor is not as constant as it is in dataset two, hence its separation from BSA is of more significance than in dataset two alone. Given the fairly constant values of O^-SA, a model of similar quality can use an indicator variable, defined as one for anionic and zero for non-ionic surfactants, in place of O^-SA ($R^2 = 0.815$, RMSE = 0.57). Although these six-parameter models are less accurate than separate models, the ability to model CMC simultaneously for charged and uncharged surfactants is a unique feature of this method.

In order to establish the predictive capability of this method, 25% of the data points were randomly removed to create a test set while the remaining 75% were remodelled; the equation generated from this regression used to predict the CMC values of the test set. This process was repeated a further three times to include all molecules in at least one test set. The results show that the models are capable of accurate prediction, with $R^2 = 0.818$ and RMSE = 0.550 when averaged overall four test sets.

Dataset 3: structurally diverse drug molecules

Data set 3 represents the most challenging of all three datsets as it contains a wider range of functional groups and molecular structures than any other model of CMC. Analysis of ASA and BSA (and *A* and *B*) for this dataset showed that the two descriptors correlate with an accuracy of about 86%. This high correlation means that if both descriptors were to be included in the same model, errors would be generated and interpretability of the model and predicted values would be unreliable. It should be noted that for all previous models low correlations between ASA and BSA (and *A* and *B*) were found. Stepwise multiple linear regressions of the five scaled surface area descriptors revealed that BSA was insignificant and highly accurate models could be made without the inclusion of BSA. Similar conclusions were reached for *B* in LFER models. Hence BSA and B are omitted from all reported models.

The results in Table 1 reveal again that PSA and TSA alone cannot model CMC as well as split four-parameter models. Scaling of the ASA descriptor yields only a 3% increase in R^2 and a 0.067 decrease in RMSE, but a notable increase of 45% is seen in $R_{cv}{}^2$. The statistics of the LFER model are almost identical with those for the four parameter scaled surface area model except in $R_{cv}{}^2$ where a difference of 75% is reported. The difference in $R_{cv}{}^2$ between the LFER and four-parameter scaled surface area model is due entirely to the inability of the LFER method to predict the value for actinomycin D when it is omitted during the cross-validation procedure, whereas the four-parameter surface area model predicts the CMC value of actinomycin D with reasonable accuracy. It should be noted that the structural diversity of this dataset is much higher that that of the previous sets, with 32 of our 46 defined fragments being employed in the assignment of scaling factors.

DISCUSSION

The significance of each descriptor in a model is given by its *t*-ratio, rather than its coefficient, so these are given in Tables 2 and 3. The pattern in *t*-ratios is fairly constant across models, with the molecular size descriptors giving

Table 2. *t*-Ratios for best surface area models

	Ionic	Non-ionic	Ionic/ non-ionic	Structurally diverse
Intercept	6.36	5.79	2.51	2.56
Area	−25.97	−22.87	−19.84	−4.45
Acid	−3.70	−19.07	−14.63	−7.18
Base	6.18	20.04	16.13	N/A
Hal	−2.83	−11.66	−8.64	−2.35
Ben	1.26	−0.24	0.62	−3.21
O^-	−2.04	N/A	0.03	N/A

Table 3. *t*-Ratios for Abraham models

	Ionic	Non-ionic	Structurally diverse
Intercept	2.85	4.36	1.65
E	−1.31	−4.73	−5.45
S	2.54	7.84	4.42
A	−3.65	−8.52	−7.28
B	1.66	4.86	N/A
V	−24.92	−19.31	−1.55

large negative values for all four datasets, indicating that larger molecules will form micelles at lower concentrations. This relationship is well established and has been stated in previous studies[34] of CMC, e.g. CMC decreases by half for every methylene added to the chain for ionic surfactants.

The molecular size terms are the most significant in all models except for dataset 3. The drop in significance of the size terms for dataset 3 is due to the complex 3D structures of the surfactants. The surfactants in datasets 1 and 2 can be predominantly split into their hydrophobic tail and hydrophilic head components, with the hydrophobic regions being mainly straight hydrocarbon chains. These can intertwine easily during micelle formation owing to their flexibility, making the process of surfactant–surfactant interaction on micellization fairly constant over all surfactants. This simple intertwining is not possible for many of the molecules in dataset 3 such as thioridazine and actinomycin D. Hence the size term for dataset 3 is forced to account for both the enthalpic and entropic factors that are needed to create a cavity in the solvent and for the surfactants and the self-association properties upon micelle formation.

Hydrogen bond acidity (*A* and ASA) and basicity (*B* and BSA) descriptors are also fairly constant in their *t*-ratio values throughout, with hydrogen bond acidity terms giving negative values and hydrogen bond basicity terms giving positive values. This result means that stronger hydrogen bonding acidic surfactants will form micelles at lower concentrations than weaker hydrogen bond acidic surfactants, whereas increasing the hydrogen bond basicity of a surfactant acts to raise its CMC.

Further insight into the role of the hydrogen bonding descriptors can be gained by comparing coefficient values from the LFER logCMC models with those for Abraham's model of aqueous solubility ($\log S_w$).[35] This comparison allows us to infer how proportionately the descriptors are representing the ability of the surfactant to interact with water and interact with themselves during aggregation. The LFER for $\log S_w$ gives large positive coefficents for *A* and *B* of 0.65 and 3.39, respectively. The LFER models of CMC also show positive values for *B*, indicating that surfactants with a larger hydrogen bond basicity can interact with water favourably, thus reducing their ability to form micelles and raising CMC.

The hydrogen bond acidity descriptor has a positive coefficient in the $\log S_w$ model, but a large negative coefficient in models of CMC. The difference in these coefficient values indicates that *A* is predominantly describing self-association effects of the surfactants and not surfactant–water interactions. It is not surprising that of the two descriptors, *A* and *B*, it is *A* that contains the information for self-association. Using the definitions of H-bond acidity and basicity stated in this study, it is possible for a molecule to be only a hydrogen bond base, i.e. contain no hydrogen attached to oxygen and nitrogen, whereas it is impossible for a surfactant to display only H-bond acidic properties. Hence any molecule with hydrogen bond acid groups will also contain hydrogen bond basic groups, giving rise to strong self-association interactions. The surface area descriptor BenSA is not highly significant in any of the models of CMC, nor is its value constant throughout all systems. For datasets 1 and 3, negative values of BenSA are displayed, as expected since it is known that the addition of one phenyl group is roughly equivalent in its effects on CMC to three methylene groups. The positive value of BenSA for dataset 2 is perhaps due to the lack of structural diversity here, since all phenyl rings are attached to an electron-withdrawing SO_3^- group.

HalSA's *t*-ratios are relatively small and negative throughout all models of CMC. Such negative values can be easily attributed the fact that halogens are hydrophobic in nature.

The *E* and *S* descriptors of the LFER approach are again easily interpreted by comparison with the LFER for $\log S_w$. The coefficient for the *S* descriptor is positive in both the CMC models and $\log S_w$ indicating that more polar surfactants will associate more preferentially with water and thus raise the CMC. The coefficient for *E* is negative in both equations, implying that surfactants with a high density of π- and n-electron pairs would rather interact with each other than with water, presumably through dispersion forces.

CONCLUSIONS

Models have been developed for the prediction of CMC for anionic and non-ionic surfactants, using the LFER approach of Abraham and surface area method of

Saunders *et al*. The models produced from these datasets were less accurate, but are more generally applicable, than those produced from previous studies. Furthermore, they allow detailed physical analysis, giving an insight into the factors determining CMC. This analysis was possible as the same descriptors were used thoughout and not chosen for each set from a larger pool of descriptors. Thus a model was created that combined molecules from the ionic and non-ionic datasets using a modified version of the surface area approach. A reasonable correlation of observed vs calculated CMC values was seen for this model, although the statistics are slightly worse than for separate models of neutral and ionic surfactants. The predictive capability of this model was confirmed by the construction of training and test sets, which showed that logCMC can be predicted to 0.55 log units.

As the structural diversity included in these surfactant models was very narrow, a model was made that included drug molecules which included a wide range of functional groups and molecular structures. The best model produced for this set was the scaled surface area model, which had an R^2 value of 0.91 and an RMSE of 0.42. This model was also examined to give us information about the micellization process and the findings were in agreement with those of other models and more importantly are physically valid.

Acknowledgement

This work was sponsored by the UK Engineering and Physical Science Research Council, GR/N20638.

REFERENCES

1. Rosen MJ. *Surfactants and Interfacial Phenomena*. Wiley: New York, 1989; 110–111.
2. Rosen MJ. *J. Colloid Interface Sci*. 1976; **56**: 320–328.
3. Meguro K, Takasawa Y, Kawahashi N, Tabata Y, Ueno M. *J. Colloid Interface Sci.* 1981; **83**: 50–56.
4. Ravey JC, Fherbi A, Stebe MJ. Progress in Colloidal and Polymer Science. *J. Prog. Colloid Polym. Sci.* 1988; **76**: 234–241.
5. Beecher P. *Dispers. Sci. Technol.* 1984; **5**: 81–87.
6. Wang Z, Li G, Zhang X, Wang R, Lou A. *Colloids Surf.* 2002; **197**: 37–45.
7. Kier LB, Hall LM. *Molecular Connectivity in Chemistry and Drug Research.* Academic Press: New York, 1986.
8. Huibers PDT, Lobanov VS, Katritzky AR, Shah DO, Karelson M. *Langmuir* 1996; **12**: 1462–1470.
9. Katritzky AR, Ignatchenko ES, Barcock RA, Karelson M. QSPR *Chem. Soc. Rev.* 1995; **24**: 279–287.
10. Stankevich MI, Stankevich IV, Zefirov NS. *Russ. Chem. Rev.* 1988; **57**: 191–208.
11. Huibers PDT, Lobanov VS, Katritzky AR, Shah DO, Karelson M. *J. Colloid Interface Sci.* 1997; **187**: 113–120.
12. Abraham MH. *Chem. Soc. Rev.* 1993; **22**: 73–83.
13. Abraham MH, Chadha HS, Martins F, Mitchell RC, Bradbury MW, Gratton JA. *Pestic. Sci*. 1999; **55**: 78–88.
14. Abraham MH, McGowan JC. *Chromatographia* 1987; **23**: 243–246.
15. Platts JA, Butina D, Abraham MH, Hersey A. *J. Chem. Inf. Comput. Sci*. 1999; **39**: 835–845.
16. Abraham MH, Chadha HS, Whiting GS, Mitchell RC. *J. Pharm. Sci.* 1994; **83**: 1085–1100.
17. Saunders RA, Platts JA. *J. Phys. Org. Chem.* 2001; **14**: 612–617.
18. Platts JA, Abraham MH, Zhao YH, Hersey A, Ijaz L, Butina D. *Eur. J. Med. Chem.* 2001; **36**: 719–730.
19. Saunders RA, Platts JA. *New J Chem.* 2004; **28**: 166–172.
20. Clark DE. *J. Pharm. Sci*. 1999; **88**: 807–823.
21. McCracken RO, Lipkowitz KB. *J. Parasitol.* 1990; **76**: 180–185.
22. Clark DE. *J. Pharm. Sci*. 1999; **88**: 815–821.
23. van deWaterbeemd H, Camenisch G, Folkers G, Raevsky OA. *Quant. Struct.–Act. Relat*. 1996; **15**: 480–490.
24. Veber DF, Johnson SR, Cheng H, Smith BR, Ward KW, Kopple KD. *J. Med. Chem*. 2002; **45**: 2615–2623.
25. Stenburg P, Norinder U, Luthman K, Artursson P. *J. Med Chem*. 2001; **44**: 1927–1937.
26. Winiwarter S, Ax F, Lennernas H, Hallberg A, Pettersson C, Karlen A. *J. Mol. Graph. Modell.* 2003; **21**: 273–287.
27. Schreier S, Malheiros SVP, De Paula E. *Biochim. Biophys. Acta* 2000; **1508**: 210–234.
28. Weininger DJ. *J. Chem. Inf. Comput. Sci*. 1988; **28**: 31–36.
29. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V. *J. Chem. Inf. Comput. Sci*. 1996; **36**: 1030–1037.
30. HyperChem 6, Hypercube Inc., 2000, Gainesville, FL.
31. Dodd LR, Theodorou DN. *Mol. Phys.* 1991; **72**: 1313–1324.
32. JMP, SAS Software, 2000, Cary, NC.
33. Huque FTT, Platts PA. *Org. Biomol. Chem.* 2003; **1**: 1419–1424.
34. Attwood D, Florence AT. *Surfactant Systems. Their Chemistry, Pharmacy and Biology*. Chapman and Hall: London, 1983.
35. Abraham MH, Joelle LE. *J. Pharm. Sci.* 1999; **88**: 868–880.