Verlag: New York, 1990; pp 254-263. (c) Martin, Y. C. Beyond Graphics: ALADDIN, A Computer Tool for Drug Design. In *Frontiers in Drug Research*; Jensen, B., Jorgensen, F. S., Kofod, H. Eds.; Alfred Benzon Symposium 28; Munksgaard: Copenhagen, 1990; pp 222-229.
(14) (a) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312-316. (b) Davies, E. K.; Upton, R. M. In *Online Information 90*, 14th International Online Information Meeting Proceedings; Raitt, D. I., Ed.; Learned Information: Oxford, 1990; p 129.
(15) Prous, J. R. *Drug Data Report*, Vol. 11; Science Publications: Barcelona, 1989.
(16) Allen, F. H.; Kennard, O.; Watson, D. G.; Brammer, L.; Orpen, A. G.; Taylor, R. Tables and Bond Lengths Determined by X-ray and Neutron Diffraction. *J. Chem. Soc. Perkin Trans. 2* **1987**, S1-S19.
(17) Rusinko, A.; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; Venkataraghavan, R. Using CONCORD To Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251-255.
(18) Hangauer, D. G. In *Computer-Aided Drug Design, Methods and Applications*; Perun, T. J., Propst, C. L., Eds.; Marcel Dekker, Inc.: New York, 1989; pp 253-295.
(19) Saunders, M. R.; Tute, M. S.; Webb, G. A. A Theoretical Study of Angiotensin-Converting Enzyme Inhibitors. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 133-142.
(20) Andrews, P. R.; Carson, J. M.; Caselli, A.; Spark, M. J.; Woods, R. Conformational Analysis and Active Site Modeling of Angiotensin-Converting Enzyme Inhibitors. *J. Med. Chem.* **1985**, *28*, 393-399.
(21) Petrillo, E. W., Jr.; Ondetti, M. A. Angiotensin-Converting Enzyme Inhibitors: Medicinal Chemistry and Biological Actions. *Med. Res. Rev.* **1982**, *2*, 1-41.
(22) Mayer, D.; Naylor, C. B.; Motoc, I.; Marshall, G. R. A Unique Geometry of the Active Site of Angiotensin-Converting Enzyme Consistent with Structure-Activity Studies. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 3-16.
(23) Marshall, G. R.; Motoc, I. In *Molecular Graphics and Drug Design*; Burgen, A. S. V., Roberts, G. C. K., Tute, M. S., Eds.; Elsevier Science Publishers: Amsterdam, 1986; p 115.
(24) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. In *Computer Associated Drug Design*; ACS Symposium Series 112; American Chemical Society, Washington, DC, 1979; p 205.
(25) Lloyd, E. J.; Andrews, P. R. A Common Structural Model for Central Nervous System Drugs and Their Receptors. *J. Med. Chem.* **1986**, *29*, 453.

Computer-Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds

DAVID T. STANTON† and PETER C. JURS*
Chemistry Department, 152 Davey Laboratory, Penn State University, University Park, Pennsylvania 16802

Received July 15, 1991

Computer-assisted methods are applied to the study of the relationship between molecular structure and observed surface tension of small organic alkanes, alkyl esters, and alkyl alcohols. Features of these molecules are encoded using a wide variety of topologic, geometric, and electronic descriptors. The simple correlation between these descriptors and observed surface tension values is examined to gain insight as to which molecular features most influence the observed surface tension. Multivariate linear regression models for each functional group class are also examined. The results of the examination of both the simple correlations and the regression models suggest that molecular surface area is an important feature. The results also show that many descriptors provide surface area information which is specific to particular portions of the molecule, and that this information provides better results in modeling surface tension than the van der Waals or solvent-accessible surface area. Finally, a multiple linear regression model is developed for a combined set of alkanes, alkyl esters, and alkyl alcohols which yields good results for predicting the surface tension of similar compounds.

INTRODUCTION

Surface tension, like normal boiling point or chromatographic retention, is a physical property which is a function of molecular structure. However, very little has been published in the chemical literature concerning the relationship between the structure of a molecule and the observed surface tension at a given temperature. The purpose of the research described here was twofold. The first goal was to establish that surface tension is a physical property that can be studied in the same fasion as other properties using quantitative structure-property relationship (QSPR) techniques and the tools available in the ADAPT software system.[1,2] The second goal of this work was to determine if such a study could shed light on the structure-property relationship involving surface tension.

The reasons for studying the relationship between molecular structure and a physical property such as surface tension are very similar to the reasons for studying normal boiling points or any other property. The material of interest may be in short supply or the experimental procedure itself may be too time consuming or expensive to be performed for more than just a few compounds. However, with the aid of a carefully developed predictive model, the values for the property of interest can be quickly and accurately estimated. But among the usual advantages which can be obtained from the modeling of a given physical property, the ability to learn something about how the structural features of a molecule can affect that property is possibly the most important. This is especially true in the case of surface tension. There have been many generalized statements made which associate polar and hydrogen-bonding intermolecular interactions with increased surface tension of pure liquids.[3,4] Surface tension has also been noted to increase as a function of molecular weight for a set of congeners.[5] However, very little more has been published in the chemical literature concerning the structure-property relationship for surface tension of pure organic compounds. Therefore, it was of interest to employ QSPR techniques to expand our understanding in this area.

In a previous paper, a brief study of the relationship between molecular structure and observed surface tensions for a small and diverse set of organic compounds was reported.[6] The results of that study suggested that it was possible to model surface tension. However, the results also suggested that additional work was necessary to improve the accuracy of the resulting models. Part of the problem associated with the accuracy of the model developed in that study involved the small size and wide diversity of the compounds in the dataset involved. In order to study the structure-property relationship for surface tension effectively, it was necessary to select a dataset of reasonable size and minimum diversity. In this way

† Present address: Norwich Eaton Pharmaceuticals, Inc., P.O. Box 191, Norwich, NY 13815.

0095-2338/92/1632-0109$03.00/0

it would be possible to determine how the values of individual molecular descriptors and observed surface tension values change with small changes in the structure. The quality of the experimental data is also important. Without good quality data, it is more difficult to correlate the true variation of the property with changes in the structure.

METHODOLOGY

All the computations done in this study were performed on a Sun 4/110 workstation running the ADAPT software system under the UNIX operating system. The general procedure for the development of regression equations from structural descriptors using ADAPT has been outlined previously.[2] Along with the structural descriptors which have been used in similar studies, the set of descriptors termed Charged Partial Surface Area (CPSA) descriptors were calculated.[6] In addition, a new set of descriptors were calculated which are similar to the CPSA parameters, but which are designed to encode information specific to molecular features capable of participating in hydrogen-bonding intermolecular interactions. Descriptor analysis and subsequent regression analysis was carried out as previously described. Models produced in this study were validated by examining the variance inflation factors for indications of high collinearity, and the jackknifed residuals were examined for evidence of observations that overly influence the calculation of the regression function.

The experimental data for this study was taken from a critically reviewed collection assembled by Jasper.[7] In order to examine the differences between models for different functional groups, sets of compounds classified as alkanes, aliphatic esters, and aliphatic alcohols were selected. The alkane dataset consisted of 95 compounds with sizes ranging from 5 to 16 carbons and included both acyclic and cyclic compounds. The ester dataset contained 56 observations of sizes ranging from 2 to 14 carbons. The alcohol dataset contained 35 observations of sizes ranging from 1 to 14 carbons. Neither the ester dataset nor the alcohol dataset contained cyclic compounds or more than one functional group. The experimental surface tension values for all the compounds were determined using the capillary rise method at 30 °C in either air or nitrogen. The uncertainty of the experimental measurements was reported to be ±0.1 dyn/cm. The identity of the compounds involved and their observed surface tension values are given in Table I which is available as supplementary material. (See paragraph at end of paper regarding supplementary material.)

RESULTS AND DISCUSSION

In order to study thoroughly the structure–property relationship for surface tension for the datasets chosen, the study was divided into three stages. The first stage involved the examination of the simple correlation of a wide variety of molecular descriptors with observed surface tension values. The second stage involved the development of models which correlated a number of descriptors with observed surface tensions for each dataset. These models were then compared and contrasted to determine how they differed and what could be learned from the differences observed. The final stage involved combining observations from each of the individual datasets into a single dataset for the purpose of developing a predictive model with good accuracy. The quality of the resulting predictive model would then be demonstrated using an external prediction set.

Examination of Simple Correlations. As is outlined above, the first step of this study involved the calculation of the pairwise correlations between a wide variety of molecular descriptors and surface tension for the compounds in each dataset. A total of 145 different molecular descriptors were examined. The hydrogen-bonding descriptors were considered only for the alcohol dataset.

Table II. Results of Determination of Correlation between Single Molecular Descriptors and Observed Surface Tension Values for Alkane, Ester, and Alcohol Datasets

descriptor	correlation coefficient
(A) Alkane/Cycloalkane Dataset	
(1) count of paths of length 1–45/no. of atoms	0.881
(2) 4th order path molecular connectivity[a]	0.837
(3) molecular ID/no. of atoms in molecule[b]	0.833
(4) count of single bonds	0.821
(5) molecular weight	0.695
(B) Ester Dataset	
(1) valence corrected 3rd order path molecular connectivity[a]	0.889
(2) valence corrected 5th orderpath molecular connectivity[a]	0.864
(3) molecular shape index ($^2\kappa$)[c]	0.832
(4) 1st order molecular connectivity[a]	0.816
(5) molecular weight	0.797
(C) Alcohol Dataset	
(1) molecular shape index, $^2\kappa$, corrected for atom and bond types[c]	0.899
(2) weighted charged partial surface area (WPSA-3)[d]	0.847
(3) molecular ID/no. of atoms in molecule[b]	0.831
(4) partial positive surface area-3 (PPSA-3)[d]	0.828
(5) molecular weight	0.786

[a] See Kier and Hall.[8] [b] See Randić.[9] [c] See Kier.[10] [d] See Stanton and Jurs.[6]

The correlation data was examined and the four descriptors with the highest correlation coefficients were selected. The four descriptors selected for each dataset are listed in Table II. Examination of the descriptors that are highly correlated indicate a trend toward the parameter types which encode polar intermolecular interactions as one proceeds from the alkane dataset to the alcohol dataset. The descriptors which are highly correlated with surface tension for the alkane dataset are all topologic in nature. The highly correlated descriptors for the ester dataset are also topologic in nature. However, descriptors involving longer paths become more important, and a molecular shape parameter ($^2\kappa$) has been included.[10] The molecular shape descriptor encodes information concerning the density (distance of all atoms to a central point) of the molecule. Finally, two CPSA descriptors are included in the list for the alcohol dataset. The CPSA descriptors have been found to be important when interactions of a polar nature occur between molecules. It was expected that polar interactions, and possibly hydrogen-bonding interactions, would influence the observed surface tensions for the alcohol dataset. However, none of the hydrogen-bonding specific descriptors were very highly correlated with the observed surface tension values.

Molecular weight has been listed for each dataset. This is because it had been reported that molecular weight was correlated with observed surface tensions for a series of congeners.[5] While a reasonable correlation exists between molecular weight and observed surface tension in each dataset, many of the other structure-based descriptors available in ADAPT yield higher correlations.

Not all the parameters examined in this portion of the study were linearly related to surface tension. For each of the 145 available descriptors, a plot was made of the correlation of observed surface tension and the descriptor in question. The plots were then examined for any sign of a pattern or nonlinear relationship. If any pattern was observed, simple mathematical transforms of the descriptor values were applied to the descriptor in order to obtain a more linear relationship. For example, the descriptor ALLP-2 (count of paths of lengths

1–45 divided by the number of atoms in the molecule)[11] was correlated to 0.881 with the observed surface tension for the 95-observation alkane dataset. By transforming the descriptor using a log (base 10) function, this correlation improved to 0.920.

There were no nonlinear relationships observed for the ester dataset. Several descriptors showed a nonlinear relationship with surface tension for the alcohol dataset. The most notable was the first descriptor (the $^2\kappa$ shape index) from Table II for the alcohol dataset. In this case also, a logarithm (base 10) function was found to give a more linear correlation with surface tension. The correlation of surface tension and the transformed variable in this case was 0.953. The third descriptor for the alcohol dataset (molecular ID/no. of atoms) in Table II also yielded an improved correlation (r = 0.877) when the log (base 10) transform was applied.

In all cases where nonlinear relationships were observed, a log (base 10) transform of the descriptor in question gave the most linear correlation. It is not clear from any of the chemical literature if the logarithmic relationship is significant with respect to surface tension for these types of compounds. It was also observed that all the descriptors for which such transforms were useful were derived from the topological representation of a molecule.

An interesting comparison can be obtained by examining how the $^1\chi$ (first-order molecular connectivity descriptor)[8] and surface tension change with changing structure. The $^1\chi$ parameter encodes the degree of branching in a molecule, and the value of the descriptor decreases as the molecule becomes more branched. For both the alkane and alcohol datasets, $^1\chi$ is positively correlated with surface tension (correlation coefficients are 0.773 and 0.819, respectively), and both yield a slight nonlinear relationship similar to that observed previously. If only the acylic alkanes are considered, the correlation between $^1\chi$ and surface tension is much stronger (r = 0.890) and a more pronounced logarithmic relationship is noted. Thus, it can be seen that as the structure becomes more branched the surface tension decreases. By taking the log (base 10) of $^1\chi$ for the alcohol dataset, the correlation with surface tension is improved (r = 0.834). The improvement is more pronounced with the acyclic alkanes where the correlation coefficient is increased to 0.912.

The observed logarithmic relationships suggest that the shape of the molecule is an important factor when the molecule is small. Highly branched structures will be more ball-shaped, and the overall surface area of the molecule will be decreased. This will decrease the interactions between molecules and therefore will reduce the observed surface tension. As larger molecules are considered, small changes in branching will have a reduced effect on the overall surface area of the molecule. This in turn will produce a smaller change in the observed surface area. The effect is somewhat reduced for the alcohol dataset, probably because of the strong polar interactions involving the hydroxyl functional group.

In order to test the idea that $^1\chi$ is encoding surface area information, a comparison of solvent-accessible surface area and surface tension was performed using the acyclic alkanes. The solvent-accessible surface area was calculated using the method of Pearlman.[12] The correlation of the solvent-accessible surrace area and surface tension values for the acyclic alkanes was 0.747, while the correlation between surface area and $^1\chi$ for the same compounds was 0.954. The relationship between surface area and surface tension appeared to be logarithmic, as was the relationship between $^1\chi$ and surface tension.

After calculating the log (base 10) for the surface area values, the correlation with surface tension improved to 0.771. These results support the idea that $^1\chi$ is encoding information concerning molecular surface area. This observation is supported by similar evidence reported by Kier and Hall.[13]

Table III. Details of Correlation Model Developed for the 95-Observation Alkane Dataset

R^2 = 0.978, s = 0.4 dyn/cm, N = 95

descriptor	regression coeff.	std. dev. of coeff.
(1) 2nd order molecular connectivity	−1.073	0.1124
(2) 5th order path molecular connectivity	−0.8873	0.2519
(3) 4th order path/cluster mol. connectivity	1.059	7.287×10^{-2}
(4) count of all paths of length 0–45	2.743×10^{-2}	2.957×10^{-3}
(5) molecular ID/no. of atoms	36.18	1.398
(6) relative negative charge (RNCG)	−56.03	3.024
intercept	−35.73	2.567

Several things have been learned through the examination of the pairwise correlations between structural descriptors and observed surface tension. Important individual descriptors, and therefore, important molecular features have been identified. A relationship between molecular surface area and surface tension has been suggested. The relationship was detected using topological descriptors which are highly correlated with surface area descriptors, but which are more easily calculated and are insensitive to changes in the geometry of the molecules. The idea that molecular surface area is correlated with observed surface tension in the bulk phase is appealing from a physical perspective. As the surface area of the molecules increases, the interactions between molecules (due to various van der Waals forces) will also increase, which would lead to higher observed surface tension values. The logarithmic relationship observed between experimental surface tension values and many of the structural descriptors also has a physical interpretation. Adding a single carbon atom to a small molecule will increase the molecular surface area by a greater amount than adding a carbon atom to a much larger molecule. The incremental increase of surface area diminishes as the molecules become large, and this effect will probably be masked as the influence of branching becomes important.

Thus, we find that much can be learned from a careful examination of individual structural descriptors. The next step is to determine if several descriptors can be combined to yield additional information.

Dataset-Specific Model Development. For each dataset, all 145 structural parameters were subjected to descriptor analysis. Once a subset of descriptors had been identified, models were developed using multiple linear regression analysis methods. Descriptors were selected in regression based on their statistical significance, but preference was given to descriptors which had a sound physical interpretation. Also, because the goal was to develop the best correlation model possible, the size of the model was only limited by the maximum allowable number of descriptors and the partial-F values for the descriptors involved. For the purposes of comparing datasets on opposite ends of the polarity scale, only the results for the alkanes and the aliphatic alcohols will be considered in this section.

Alkane Dataset Modeling. The alkane dataset was considered first. A model of reasonable quality was obtained. The details of the model are given in Table III. The scatter plot of the fitted and observed surface tension values is shown in Figure 1. The model employs six descriptors, five of which are topologic in nature. The sixth descriptor (relative negative charge, RNCG) is a CPSA descriptor which encodes electronic information. The standard deviation of regression (s) for the model is 0.4 dyn/cm, which represents an error of 1.9% based on the mean surface tension for the alkane dataset. This result is much improved over the results obtained in the past.[6] The regression fit error is also similar in magnitude to the exper-

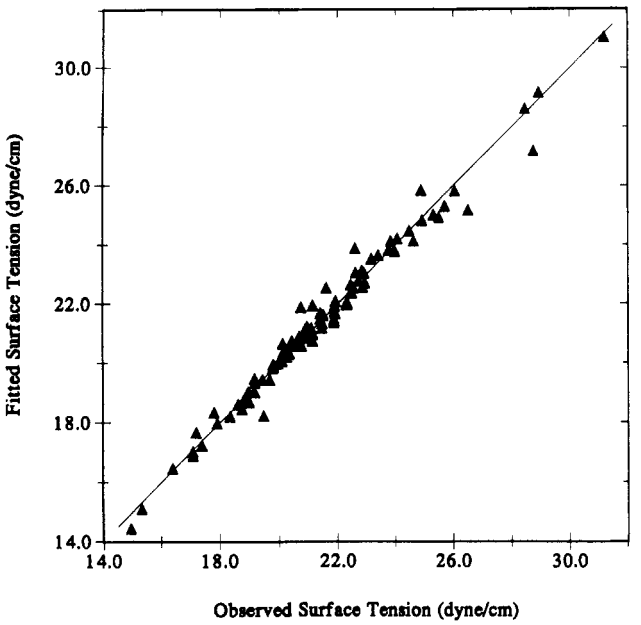

Figure 1. Scatter plot of the fitted and observed surface tension values for the alkane/cycloalkane class-specific dataset.

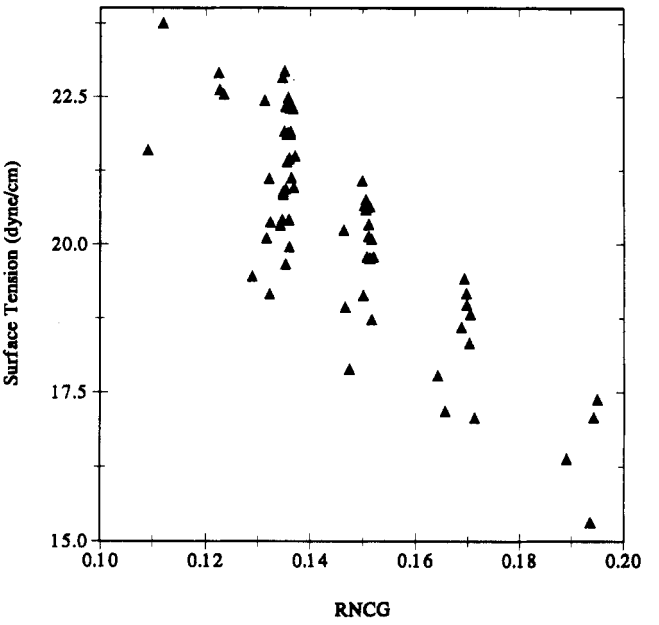

Figure 2. Scatter plots showing the correlation of the CPSA descriptor RNCG with the observed surface tension values for 76 acyclic alkanes.

imental error for the observed surface tension values.

Examination of the descriptors involved in the model yields several interesting observations. Only one of the descriptors listed in Table II which were found to be highly correlated to surface tension was found to be included in the model. This suggests that while a given descriptor is highly correlated with the dependent variable, it is not necessarily as important when taken in combination with other descriptors. In general practice, regression is begun by using the descriptor which is most highly correlated to the dependent variable. Later, as other descriptors are added to the model, the significance of the first descriptor often becomes so low that it must be dropped from the model. Therefore, one can assume that the combination of descriptors remaining in the model encodes the same information as the first descriptor in addition to other important information.

The five topological descriptors in the model present no surprises when considering the alkane dataset. Such descriptors bring size, shape, and branching information to the model. It is suggested by the observations made in the preceding section that this information is related to the molecular surface area and that the observed surface tension values are ultimately related to molecular surface area. The need for more than one descriptor to encode molecular surface area may be because the information provided by each descriptor may be incomplete in itself or that the different descriptors provide molecular surface area information for different portions of a molecule.

Probably the most interesting descriptor in the alkane model is the CPSA descriptor RNCG. It was not clear why an electronic descriptor would be significant in a model based on an alkane dataset. However, RNCG is the second most important descriptor in the model in terms of its contribution to the final estimated surface tension value. The contribution of each descriptor was determined by taking the mean values for each descriptor and combining them with the coefficients of the original model to estimate the mean surface tension value. The importance of each descriptor is then related to the fraction of the estimated surface tension which is contributed by that descriptor. For the alkane model, the most important descriptor is the ratio of the molecular ID to the number of atoms in the molecule (descriptor 5 in the model), and RNCG contributes the second largest amount to the final surface tension value.

In an attempt to visualize what information the RNCG descriptor may be bringing to the model, the observed surface tension values were plotted against RNCG for the acyclic alkanes. This plot is shown in Figure 2. The acyclic alkanes were considered in this example in order to simplify the graph. In Figure 2, a series of clusters is observed, and the general correlation for surface tension and RNCG is negative. It was determined that the clusters represent a division of the dataset based on the size (number of carbons) of the molecule. Thus, one cluster represents the C_6 isomers and another represents the C_7 isomers. However, there appears to be a distribution of data points within each cluster as well.

The RNCG descriptor is calculated by taking the largest negative partial atomic charge, calculated using the method of Abraham and Smith,[14] and dividing that value by the total negative charge on the molecule. The result is a value which indicates the fraction of the negative charge associated with the most negatively charged atom. For alkane molecules, the negative charges are carried by the carbon atoms while the hydrogens are the positively charged species. While differences in the negative charge among the carbons are not large, certain atoms usually possess more negative charge than others. For a normal alkane (e.g., *n*-butane), the largest negative charges reside on the terminal methyl groups. The carbon atoms of the two methyl groups will have identical partial charges, and the methylene groups in the chain will also have the same charge as each other. As the size of the molecule increases, there are more carbons over which to spread the negative charge. Thus, the terminal methyl groups possess a smaller fraction of the total negative charge. This explains the clusters observed in Figure 2.

As the branching in the molecule changes, the charge distribution also changes. This explains the distribution of the data points within each cluster. Figure 3 focuses on the cluster representing the C_9 alkane isomers. Within the cluster, a positive correlation between surface tension and RNCG is observed for the set of tetramethyl pentane isomers. For the 2,2,4,4-tetramethyl isomer, the six terminal methyl groups possess identical partial charges. The branching is spread evenly over the molecule, and the compound exhibits a lower observed surface tension. In the case of the 2,2,3,3-tetramethyl isomer, one of the terminal methyl groups (carbon-5) is isolated, and it is the most negatively charged atom. This compound exhibits a higher observed surface tension. The positive

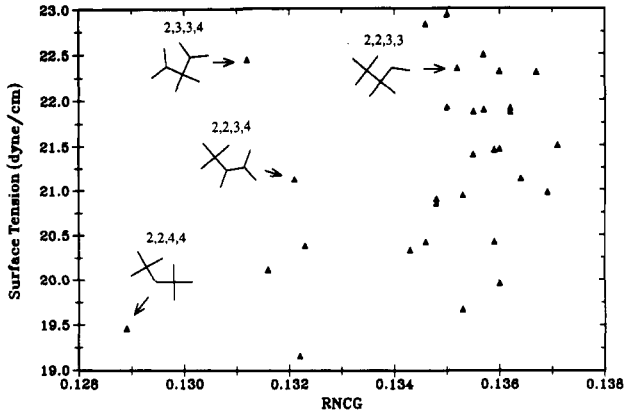

Figure 3. Correlation of the CPSA descriptor RNCG with the observed surface tension values of C_9 acyclic alkane isomers.

Table IV. Details of Correlation Model Developed for 35-Observation Alcohol Dataset

$R^2 = 0.975$, $s = 0.3$ dyn/cm, $N = 35$

descriptor	regression coeff.	std. dev. of coeff.
(1) $\log_{10}$ (molecular shape index $^2\kappa$)	5.388	0.4909
(2) valence corrected 4th order path/cluster molecular connectivity	2.585	0.3924
(3) charge separation distance	−1.592	0.1791
(4) partial negative surface area (PNSA-1)	−0.1581	2.163×10^{-2}
(5) fractional negative surf. area (FNSA-3)	87.52	15.13
(6) average surface area of H-bond acceptors (RSAA)	0.2939	4.203×10^{-2}
intercept	27.33	1.196

correlation between surface tension and descriptors which encode the degree of branching of molecules was observed previously for $^1\chi$ index. The conclusion is that the RNCG descriptor was really providing information concerning the surface area of the molecule and that the surface area of a molecule is the key structural feature which affects the observed surface tension values. The RNCG descriptor appears to provide a complex combination of information concerning both molecular size and molecular surface area. The information provided must be unique since RNCG is not highly correlated with any of the topological descriptors.

Alcohol Dataset Modeling. The process of model development was repeated using the aliphatic alcohol dataset, and the result was a six-variable model. The details of the alcohol model are given in Table IV. The scatter plot of the fitted and observed surface tension values is shown in Figure 4. The fit for the alcohol model ($R^2 = 0.975$) is similar to that observed for the alkane dataset, and the standard deviation of regression is lower ($s = 0.3$ dyn/cm).

A variety of descriptor types are involved in the alcohol model, and their interpretation should be taken in the context of the observations which have already been made for the alkane dataset model and the individual descriptor correlations. The first two descriptors in the model (descriptors 1 and 2) are topologic in nature and describe molecular size and shape. The next descriptor (descriptor 3) is the charge separation distance. This descriptor is calculated by taking the through-space (Euclidian) distance between the most negatively and most positively charged atoms which are not bonded to each other. At first glance, such a descriptor should not be important for a monofunctional alcohol. The oxygen will always be the most negatively charged atom. The most positively charged atom (excluding hydrogens) will be the carbon to which the oxygen atom is directly bonded. Since

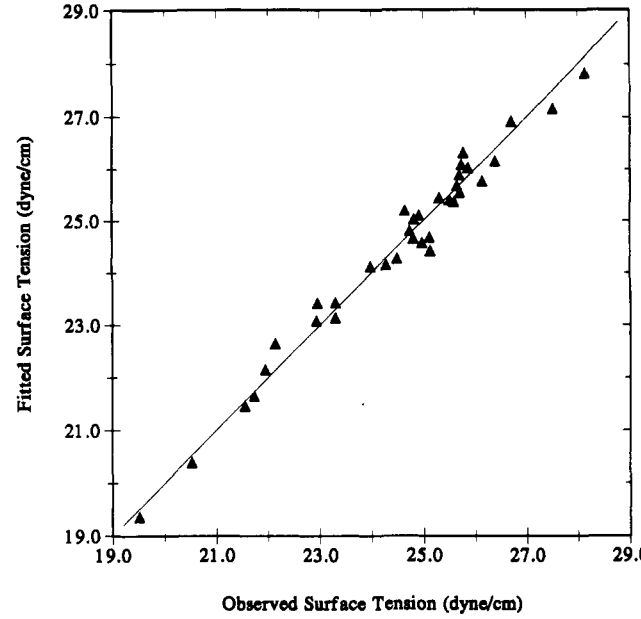

Figure 4. Scatter plot of the fitted and observed surface tension values for the aliphatic alcohol class-specific dataset.

that carbon is not considered, the carbon which is β to the oxygen will be the next most positively charged atom. However, this will depend upon where the hydroxyl group is on the carbon backbone. It may be possible that another atom several bonds away will be slightly more positive than the carbons already noted or that the conformation of the molecule around the hydroxyl group is different because of steric interactions. Thus, this descriptor is providing information on how the alcohol is constructed and the placement of the hydroxyl group within the framework of the carbon backbone.

There are two CPSA descriptors (descriptors 4 and 5) in the model, and these are both providing information about the surface area of the negatively charged atoms of the molecule. As previously noted, the carbon and oxygen atoms will carry the negative charges. However, the oxygen will generally have more exposed surface area than any single carbon atom because it is bonded to only one carbon atom and one hydrogen atom. Therefore, the CPSA descriptors are providing information concerning the environment about the hydroxyl group. As the hydroxyl group is moved farther toward the interior of the molecule, the surface area of the oxygen will be reduced due to the steric bulk of the surrounding atoms. Since the interactions between molecules with respect to the hydroxyl group will be polar (or specifically hydrogen bonding) in nature, the interactions will be limited if the contact surface for the oxygen is reduced due to crowding. This in turn will reduce the influence of the hydroxyl group on the observed surface tension of the molecule.

The final descriptor in the model (descriptor 6) is one of the hydrogen-bonding specific descriptors. It was also surprising at first glance that this descriptor would be important in a model for monofunctional alcohols because it represents the surface area of the oxygen in the molecule. However, as described above, the available surface area of the oxygen will change with its surroundings, and this will change its availability for interaction with other molecules and reduce its influence on observed surface tension values. That this descriptor is significant in a model which contains the two CPSA descriptors can be explained. The two CPSA descriptors include the surface area of the carbon atoms in addition to that of the oxygen. The information available from the CPSA descriptors relates to the conformation of the molecule overall, while the hydrogen-bonding descriptor relates specifically to the oxygen atom. This difference allows for the influence of the hydroxyl

Table V. Details of Multiple Linear Regression Model Developed for 146-Observation Combination Dataset

$R^2 = 0.983$, $s = 0.4$ dyn/cm, $N = 146$

descriptor	regression coeff.	std. dev. of coeff.
(1) 3rd order path molecular connectivity	1.632	0.1320
(2) molecular ID/no. of atoms	27.77	1.709
(3) molecular shape index ($^2\kappa$)	0.2367	3.238×10^{-2}
(4) average distance sum mol. connectivity[a]	1.232	0.1458
(5) Wiener no.[b]	-4.778×10^{-3}	9.835×10^{-4}
(6) count of paths 0–45/no. of atoms	0.4291	6.378×10^{-2}
(7) dipole moment	−0.7262	0.1063
(8) relative positive charge (RPCG)[c]	6.377	0.7085
(9) sum H-bond donor surf. area/mol. surf. area (RSHM)	37.78	1.892
(10) sum H-bond acceptor surf. area/mol. surface area (RSAM)	28.82	1.882
intercept	−40.13	3.143

[a] See Balaban.[15] [b] See Wiener.[16,17] [c] See Stanton and Jurs.[6]

group to be specifically recognized. Using the technique described above for determining the importance of a given descriptor in the model, it was observed that the surface area of the oxygen was the second most important parameter. This suggests that the availability of the oxygen for hydrogen-bonding interactions is very important in determining the surface tension of the compound.

From these experiments, it has been observed that it is possible to obtain good correlations between molecular structure descriptors and observed surface tension values for the compounds studied. It was also observed that a more complete understanding of the structure–property relationship for surface tension can be obtained by careful analysis of the descriptors found in the models developed for a given dataset. Since it has been found that it is possible to build models which seem to agree with a physical understanding of surface tension, the same techniques should provide a model which can be used for predictive purposes.

Development of a Predictive Model for Surface Tension. For this portion of the study, 166 compounds were selected from the three datasets studied thus far. From the 166 compounds, 20 compounds were selected at random to serve as an external prediction set, while the remaining 146 compounds were used to develop the predictive equation. The selection of the prediction set was done in such a way that the number of compounds from each dataset remaining in the training set would be roughly proportional to their overall availability. Thus, 74 compounds (50.7%) were taken from the alkane dataset, 44 compounds (30.1%) were taken from the aliphatic ester dataset, and the remaining 28 compounds (19.2%) were taken from the alcohol dataset. The development of the model was performed as described previously. Selection of the final model was based on having the largest coefficient of multiple determination (R^2) with the smallest model possible.

The model obtained for the combined dataset contained 10 descriptors and is detailed in Table V. The scatter plot for the fitted and observed surface tensions for the 146 observations is shown in Figure 5. A good fit was obtained for the dataset ($R^2 = 0.983$), and the standard deviation of regression was similar to that obtained for the individual dataset models ($s = 0.4$ dyn/cm).

Examination of the model shows that 6 of the 10 descriptors are topologic in nature, 2 are electronic in nature, and the remaining 2 are from the hydrogen-bonding set. The general types of descriptors found in this model are similar to the types found in the individual class models.

The utility of the combined dataset model was determined using the 20-observation external prediction set which had been

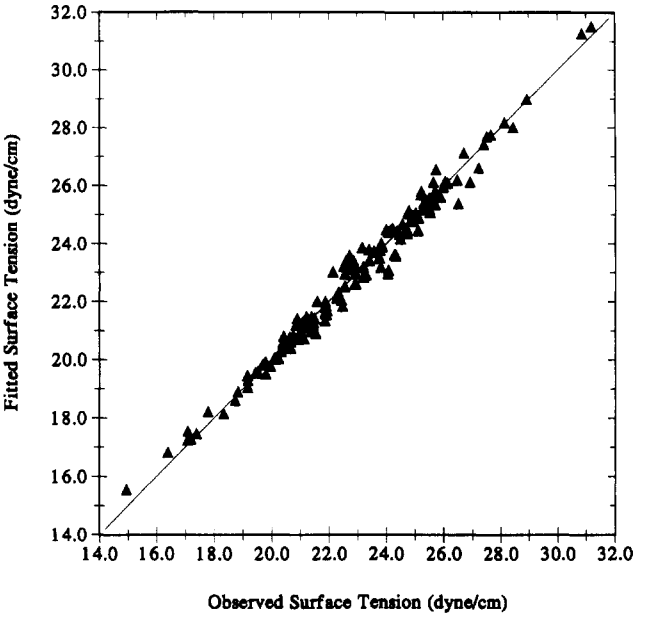

Figure 5. Scatter plot of the fitted and observed surface tension values for the alkane/ester/alcohol combination dataset.

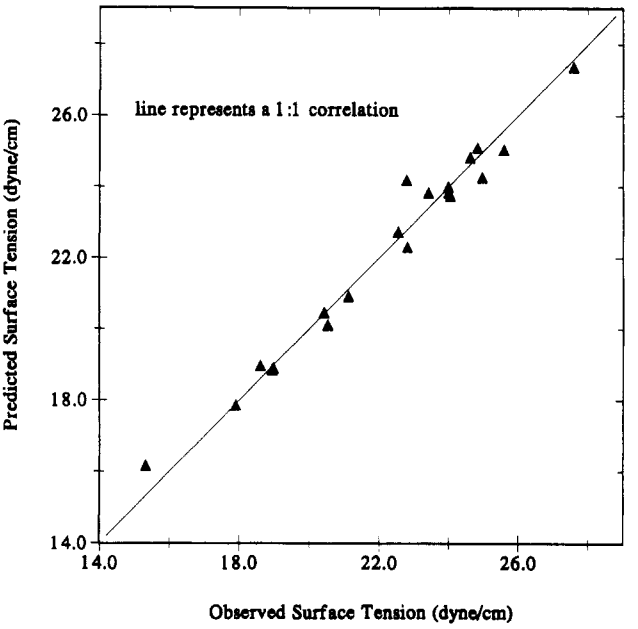

Figure 6. Scatter plot of the predicted and observed surface tension values for the 20 external combination dataset compounds.

selected prior to the beginning of the model development step. The results of the predictions are shown graphically in Figure 6. A good correlation between the predicted and observed surface tension values for the dataset ($r = 0.987$) with a total RMS error of 0.7 dyn/cm. The higher than expected RMS error value is attributed to the influence of one compound in the prediction set. The prediction error for propyl isobutyrate was −1.4 dyn/cm. The RMS error calculated on the basis of the other 19 compounds was 0.4 dyn/cm suggesting that there may be some deficiency in the model as it concerns propyl isobutyrate or that there is some sort of error associated with the observed surface tension value. However, the results for the remaining 19 compounds indicated that the model is useful and reasonably accurate. The RMS error for the prediction set (ignoring the outlier) is the same as the standard deviation of regression for the prediction model and represents a prediction error of roughly 1.8%. This is quite acceptable for an estimation method considering that the experimental error for the capillary rise method is 0.4%.

CONCLUSIONS

The results of the three stages of this study have provided a great deal of information concerning the relationship between molecular structure and the observed surface tension values for the compounds studied. Much of what has been learned is due to the availability of a wide variety of molecular structure descriptors within the ADAPT system.

In the absence of polar interactions, the most important feature of molecular structure which influences the observed surface tension appears to be molecular surface area. A number of descriptors were found which provide this type of information from various points of view. When polar interactions are possible, they do influence the observed surface tension, but the nonpolar or dispersive interactions remain important. Thus, descriptors which provide information concerning both types of interactions are required in modeling surface tension for polar compounds.

In the case of specific types of polar interactions, like hydrogen bonding, the available surface area of the acceptor group is an important factor in determining its effect on the observed surface tension. A hydrogen bond has a length equal to the diameter of the hydrogen atom being shared. If the approach of the acceptor group is hindered by the steric bulk around the donating group, the degree of hydrogen bonding for that compound will be diminished. Using such information, together with the information concerning the influence of molecular surface area, it may be possible to design a molecule with a desired level of hydrogen bonding, and thus control the observed surface tension.

While it is important to note that surface tension is related to molecular surface area, it is also important to note that the necessary information can be provided by parameters which are topologic in nature. Most calculations of molecular surface area that are based on the geometry of the molecule are approximations. Such calculations are also very time intensive and require that molecular modeling be performed before the surface area can be calculated. In contrast, the topologic descriptors, like the molecular connectivity indices, are easily calculated and are not related to geometry, but yield values which are highly correlated to the calculated surface area values. It may be necessary to combine two or more topologic descriptors to get sufficient information to estimate surface tension, but since they are easily calculated and have a physically meaningful interpretation, the number required does not present a serious problem.

Finally, it has been shown that it is possible to produce a model for the purpose of estimating surface tension for alkanes, aliphatic esters, and aliphatic alcohols. While the estimation error is larger than the experimental error, it is still quite reasonable. Such predictive equations can allow the chemist to obtain a reasonable estimate of the surface tension in a very rapid fashion. Additional work is necessary to expand the utility of such predictive models to other types of compounds. The success of that work will depend on the availability of experimental data of reasonable quality.

ACKNOWLEDGMENT

The funding for this work was provided by the Beilstein Institute. Partial funding was also provided by the National Science Foundation for the purchase of the Sun 4/110 workstation.

Supplementary Material Available: Table I giving names and experimental surface tension data for the compounds studies (5 pages). Ordering information is given on any current masthead page.

REFERENCES AND NOTES

(1) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1989**, *61*, 1328.
(2) Stanton, D. T.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.
(3) Girifalco, L. A.; Good, R. J. *J. Phys. Chem.* **1957**, *61*, 904.
(4) Masterton, W. L.; Slowinski, E. J. *Chemical Principles*; W. B. Saunders: Philadelphia, PA, 1973; p 207.
(5) LeGrand, D. G.; Gaines, G. L. *J. Colloid Interface Sci.* **1975**, *51*, 338.
(6) Stanton, D. T.; Jurs, P. C. *Anal. Chem.* **1990**, *62*, 2323.
(7) Jasper, J. J. *J. Phys. Chem. Ref. Data* **1972**, *1*, 841.
(8) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Design*; Academic Press: New York, 1976.
(9) Randić, M. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164.
(10) Kier, L. B. *Quant. Struct.–Act. Relat.* **1985**, *4*, 109.
(11) Randić, M. *Comput. Chem.* **1987**, *3*, 5.
(12) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel-Dekker: New York, 1980.
(13) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; John Wiley & Sons: New York, 1986; p 48.
(14) Abraham, R. J.; Smith, P. E. *J. Comp. Chem.* **1988**, *9*, 288–297.
(15) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
(16) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(17) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 2636.