Application of the Multiple Computer Automated Structure Evaluation Methodology to a Quantitative Structure–Activity Relationship Study of Acidity

GILLES KLOPMAN* and DAN FERCU

Department of Chemistry, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106

Received 20 December 1993; accepted 18 April 1994

ABSTRACT

We applied the Multiple Computer Automated Structure Evaluation (Multi-CASE) program to the analysis of the relationship between the structure of 2464 organic acids and their (first) pK_a values. By using the self-created expert dictionary of molecular attributes pertinent to acidity, the program could make successful *a priori* prediction of the acidity of new organic compounds. © 1994 by John Wiley & Sons, Inc.

Introduction

The determination and interpretation of the acidity of organic compounds plays an important role in physical organic chemistry, as well as in biochemistry and medicinal chemistry.[1] The biological activity of a drug depends on its ability to cross membranes between the site of administration and the site of action. The acid-base properties often determine whether the drug exists as a charged species and remains in the aqueous phase, or as an uncharged one that is more amenable to cross membranes. Indeed, weak organic acids and bases are usually more lipophilic and diffuse across the lipid regions of membranes more readily. Therefore, the ability to predict the acidity of a drug is expected to be useful in drug design. There are several articles that deal with gas-phase acidities, but only a few to our knowledge are aimed at predicting water acidity constant, pK_a, of acids.[2]

Grüber and Buss calculated the pK_a values of 182 phenols and aromatic and aliphatic carboxylic acids using a quantitative structure-activity relationship (QSAR) with four descriptors: the energy difference between the acid and its corresponding anion, the highest occupied molecular orbital

* Author to whom all correspondence should be addressed.

CCC 0192-8651 / 94 / 091041-10

(HOMO) energy of the anion, and the atomic charge densities at the hydroxylic oxygen in the acid and in its conjugate base, respectively.[3] These descriptors, obtained from semiempirical calculations from available QCPE packages, gave the best r value of .938. Ohta used the electron density of the phosphorous atom (obtained from the semiempirical PM3 method) to calculate the pK_a values (r = .918) of 36 alkylphosphonic acids.[4]

A priori estimates of pK_a's for a few organic compounds in water have been obtained from their gas-phase acidities (using *ab initio* calculations) and solvation free energies.[5,6] The major problem here is the evaluation of the effect of the solvent on the gas-phase pK_a's, which in these cases was calculated by using Monte Carlo simulations[5] or continuum dielectric methods.[6]

The aforementioned calculations are limited to a small number or a single class of compounds. In this article, we apply our Multiple Computer Automated Structure Evaluation program to study the structure-acidity relationships of a wide variety of organic compounds and drugs. The Multi-CASE program, described elsewhere,[7] is entirely computer automated and is able to select automatically the features (descriptors) associated with a specific activity from diverse training databases. In our case, the program was trained with a database of 2464 molecules of various acidities obtained from several sources.[1,8,9] Because the parameters relevant to acidity are identified by the program, they can be used to predict the acidity of new compounds and, possibly, to efficiently optimize their structure to achieve the desired acidity.

We used the program trained with the 2464-acid database to test Grüber and Buss's data set of 182 compounds. The concordance between experimental and predicted results was 98.8% and the r value .940. Using the same training set, we also tested all (214) acidic drugs available from Newton and Kluza's compilation.[10] These compounds are highly noncongeneric with respect to those from the training set; many of them have several functional groups. In this case, the concordance was 90.5% and the r value .838. When we added these molecules (see, however, Results and Discussion) to the training set and tested them again, the statistical parameters improved considerably (concordance 100% and r value .918).

The pK_a value predicted by our program can also be used as a standard descriptor in QSARs. Schultz used pK_a as a molecular descriptor in QSAR studies, along with the logarithm of the octanol/water partition coefficient (log P).[11]

The pK_a value may also be incorporated as an embedded descriptor in the apparent octanol/water partition coefficient.[12] The pH-dependent, apparent log P is based on the partition coefficients of both the ionized and un-ionized forms. Because only the un-ionized form of a drug is able to cross lipid membranes, its fraction (pH- and pK_a-dependent) along with its partition coefficient (calculated from structure[13]) can approximate the apparent partition coefficient and be used as a more appropriate descriptor in QSARs.

Methodology and Databases

MULTI-CASE APPROACH

The Multi-CASE program is an artificial intelligence system that automatically analyzes the biological or physicochemical activity of a given set (learning or training set) of compounds and identifies the structural descriptors that may be responsible for their activity or inactivity. The required input includes structural formulas and activities (pK_a values, in this case[†]) of the compounds in the database. The structural formulas are encoded using the KLN code.[14] All molecular structures are fragmented into all possible substructures (descriptors) of 2 to 10 linearly connected heavy atoms, which are labeled as active or inactive depending on whether the parent molecule is acidic. The significant fragments—those with the highest probability of being responsible for acidity (found in a binomial distribution to have at most a 5% chance of being observed if their occurrence were random)—are considered and can be used as descriptors to perform a multivariate regression analysis.

The program uses a hierarchical approach when it deals with databases consisting of compounds with different functionalities and/or mechanisms of action. The whole database is searched to identify the fragment (biophore) which can account for the activity of the largest number of compounds in the database. The compounds associated with this biophore are removed, and the program analyzes the remainder of the data until all active compounds are accounted for or no more biophores can be found. For each subset of compounds characterized by a common biophore, a local QSAR is

[†] However, we used pK_b instead of pK_a for computational convenience and to compensate for the temperature dependency of both pK_a and pK_w; $pK_w = pK_a + pK_b = 14.0$ at 25°C, 14.17 at 20°C, and 13.84 at 30°C.

performed. The fragments (modulators) in each QSAR equation either increase or decrease the activity due to the biophore. In addition to fragments, molecular physicochemical (partition coefficient,[13] water solubility,[15] molecular weight) and quantum mechanical (HMO charge densities, HOMO and LUMO coefficients, and absolute electronegativity and hardness[16]) properties[17] are included as potential descriptors in the QSAR equation.

DATABASE

The database of 2464 molecules was assembled mostly from the compilation of Kortüm et al.,[8] but other sources were also used.[9] It contains 242 aliphatic and alicyclic carboxylic acids, 200 aromatic carboxylic acids (from which 107 have the COOH group on the aromatic nucleus and 93 on the sidechain), 313 phenolic acids, 55 other acids (5 sulphonic acids and 50 phosphorus acids), and 305 special types (79 acidic OH, 18 acidic NH, 40 acidic CH, 70 heterocyclic acidic H compounds, and 98 aminoacids). An additional 1349 bases, taken from Perrin's compilation,[18] were added to this compilation as inactives to increase the knowledge base of the program.

The input for activity is the negative logarithm of the thermodynamic ionization constant, pK_a. The pK_a range of interest is between 0 (for the most acidic compounds) and 9 (for the least acidic ones). We were not interested in evaluating pK_a's beyond this range because we intend to use the pK_a prediction capability of our program mostly in drug-design studies (the lowest pK_a of a rapidly absorbed acidic drug and the highest pK_a of a similarly absorbed based drug are known to be 3 and 8, respectively[19]). Therefore, we arbitrarily assigned the aforementioned limits whenever the pK_a values exceeded them. Most pK_a values are given at 25°C, but some in the 20–30°C range or obtained at room temperature were also included to increase the training set.[‡] The breakpoint between actives and inactives was set at 7. With this breakpoint, the database consisted of 704 actives (pK_a = 6.5 or less), 55 marginals (pK_a = 6.5–7.8), and 1705 inactives (pK_a = 7.8 and higher).

Some compounds were not included in our training set due to the conditions imposed by the program. Among them are compounds containing two or more acidic groups. This was done not to confuse the computer-automated assignment of the given pK_a value to a specific molecular functionality when several exist. For example, a compound that is a carboxylic acid as well as a phenol would be taken as having the COOH group as a biophore and phenolic OH group as a modulator if the database contained more carboxylic acids than phenols, as is the case with our database, whereas the phenolic OH group would be the biophore and COOH the modulator if the database had more phenols. Compounds that include two or more identical biophores unsymmetrically displaced were also left out. For example, monomethyl- or trimethylsuccinic acids contain two COOH groups within different environments each. Geometrical isomers of unsaturated compounds were included in the database when both substituents of the double bond contained heteroatoms. In addition, conformational (meso and rac) isomers (such as 2,3-diethylsuccinic acid), cis (*Z*) isomers of unsaturated compounds (where at least one of the substituents contains no heteroatom); (e.g., isocrotonic or isocinnamic acid), all geometrical isomers of alicyclic compounds (e.g., *cis*- and *trans*-2-methylcyclohexanecarboxylic acids) were not included because the program deals only with 2D fragments at this time. All the aforementioned rules are not considered if the compound is not acidic. Inorganic molecules and compounds containing elements other than C, H, halogens, O, S, N, and P were not considered either.

The program can only handle neutral species. Thus, zwitterions, such as aminoacids, are entered into the training database in their neutral form. As input for activity, the lowest pK_a—that is, the pK_a of the (most) acidic group—was entered. For ordinarily amphoteric compounds, with the acidic pK_a larger than the basic pK_a, the input for activity was the higher value (i.e., that of the most acidic group in the neutral molecule).

For tautomers, only the most stable form was entered, usually the one with the weaker acidic proton. In the case of ascorbic acid, for example, the most stable (dienol) form is more acidic than the diketo derivative; similarly, acetaldehyde would be entered rather than the tautomeric vinyl alcohol.

Finally, it is important to mention the usefulness of the program in identifying the duplicates and the wrongly entered pK_a data from the literature (pK_b instead of pK_a or vice versa, lack of data for the most acidic site in a polyacid, pK_a of

[‡] pK_a is practically temperature insensitive for carboxylic acids and decreases linearly with temperature for phenols (see A. Albert and E. P. Serjeant, *The Determination of Ionization Constants, A Laboratory Manual*, Chapman and Hall, London, 3rd ed., 1984).

the conjugate acid of a basic site instead of pK_a of an acidic site, etc.). Therefore, it was often necessary to refer to the primary literature given in the compilations.

Results and Discussion

VALIDATION

The database is highly noncongeneric and, therefore, needs to be statistically validated. Thus the database was divided into four subsets of 616 compounds each. We ran the Multi-CASE program on the first subset and used the results to predict the acidities of the molecules of the other three subsets. Then we submitted the combined first two subsets to the program and used the results to predict the acidity of the third and fourth sets. Finally, the first three sets were used as a training set by the program and the last one as a test set (see Table I).

Table I shows that all statistical parameters have good values, which vary in the expected direction (except for specificity, which decreases insignificantly). That means that the predictive power of the program improves, although slightly, as it is acquiring more knowledge (i.e., the database it is based on is becoming larger). However, the variation of the values of these parameters is larger when going from session 1 to 2 than when going from session 2 to 3 where Phi^2 and the observed concordance no longer improve. Therefore, we can say that the knowledge power of the program fits an exponential behavior, with the largest achievement in the first step.

RESULTS OF THE MULTI-CASE ANALYSIS

The 2464-acid database was submitted to the Multi-CASE program. The program identified 22 biophores (see Table II). Some of them are well-known acidic functionalities (—COOH, $—PO_3H_2$, $—SO_3H$, phenolic —OH), others would have been harder to anticipate. As one may expect, the most important biophore is the carboxylic group (see Table II). It occurs in 522 compounds (518 actives, three marginals, and one inactive) with an average activity of 3.6 pK_a units.

The thiolocarboxylic (—COSH) and dithiocarboxylic (—CSSH) groups, encountered in the training set, are automatically selected by the program not as biophores but as expanded fragments of the COOH biophore. These fragments are defined[7] as being similar to a biophore but not existing in sufficiently large numbers in the learning set to be biophores by themselves. The similarity is related to the use of bioisosteric groups (i.e., groups with relatively similar shape and electronic properties). Once again, we emphasize that we are interested in drug-design-related studies, and the use of bioisosterism is beneficial in defining expanded fragments. In this way the program also identified the PO—OH and SO_2—OH moieties (where the CO group of the biophore is replaced by the PO or SO_2 group), therefore accounting for the acidity of phosphorous acids (phosphinic and phosphonic acids and phosphoric esters) and sulfonic acids. The nitroamino and sulfonamido groups were selected based on the assumption that NO_2 and NH are bioisosteres of CO and OH groups, respectively. The main biophore, with its expanded fragments, is found in 601 compounds, which represents about 54% of all acids in the database.

The second biophore and its corresponding expanded fragments (ethenol substituted in the second position with electron-withdrawing groups) accounts for the acidity of *o*-nitrophenols, 2-hydroquinones, tetronic acids, and 5-pyrazolones. It is present in 58 compounds, of which one is inactive and eight are marginal. The nitro-containing phe-

TABLE I.
Statistical Parameters of the Multi-CASE Predictions of the Acidity.

Set	Training / Test	Phi^2	OC, %	Sens. %	Spec. %	*r*	SD
1	616 / 1848	0.861	97.0	90.5	99.7	.925	0.907
2	1232 / 1232	0.902	97.9	93.8	99.6	.942	0.831
3	1848 / 616	0.897	97.8	94.3	99.3	.952	0.774

Phi^2 measures the accuracy of the predictions with respect to expectations from acidity randomness and equals 1 for a perfect fit. Observed concordance (OC) is the ratio of the sum of true positives and true negatives divided by the total number of predictions. Sensitivity (sens) represents the probability of an experimentally active compound to be predicted active. Specificity (spec) renders the probability of an experimentally inactive compound to be predicted inactive. The *r* value is the correlation coefficient between the experimental and predicted acidities and is 0 for no correlation and 1 for a perfect one. The SD value is the standard deviation.

TABLE II.
Multi-CASE Biophores of Acidity.

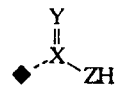
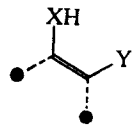
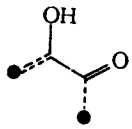
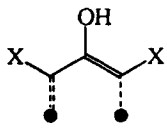
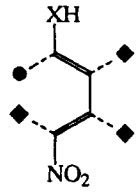
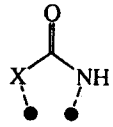
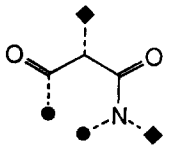
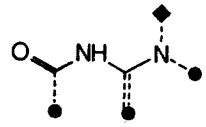
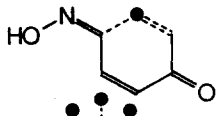
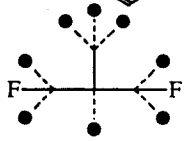
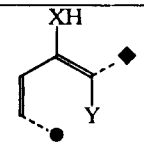
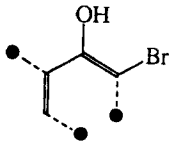
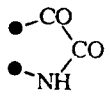
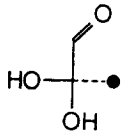
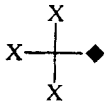
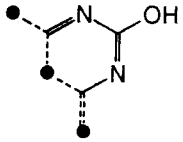
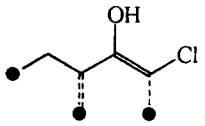
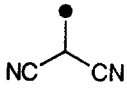
Biophore	Structure	X	Y	Z	Biophore-containing compound(s)
1		C	O	O	carboxylic acids
		C	O	S	thiolocarboxylic acids
		C	S	S	dithiocarboxylic acids
		H—P^V	O	O	phosphinic acids
		HO—P^V	O	O	phosphonic acids
		O—P^V	O	O	phosphoric esters
		N^+O^-	O	N-	nitroamines
		SO	O	N-	sulfonamides
		SO	O	O	sulfonic acids
2		O	NO_2		2-nitrophenols
		O	CO		2-hydroxyquinones tetronic acids
		N	CO		5-pyrazolones
3					hinokiols 2-hydroxyquinones
4					tetrazoles
5		F, Cl, I			2,6-dihalogenophenols
6		O			4-nitrophenols
		S			4-nitrothiophenols
7		O			5,5-disubstituted- oxazolidine-2,4-diones
		S			5,5-disubstituted- thiazolidine-2,4-diones
8					barbituric acids
9					xanthines
10					quinonemonoximes
11					polyfluoroalcohols

TABLE II.
(Continued)

Biophore	Structure	X	Y	Z	Biophore-containing compound(s)
12		S	*trans*-C_{arom}		2- and 3-substituted thiophenols
		O	*cis*-CN		2-cyanophenols
13					2-bromophenols
14					parabanic acid
15					mesoxalic dialdehyde
16		CO, NO_2, SO_2, NO_2			trisubstituted methanes trinitroethanol
17					cyameluric acid
18					2,2′-methylenebis-4,6-dichlorophenol
19					4-nitrosophenol
20					4-chlorothiophenol
21		CHO, NO_2			disubstituted methanes
22					tricyanomethane

Biophores are tabulated in the decreased order of their probabilities of relevance to acidity. • represents a nonhydrogen substituent, and ♦ stands for either a hydrogen or a nonhydrogen substituent.

nolic fragment is the main biophore and the keto derivatives are expanded substructures.

The program did not find a simple way to explain the acidity or lack of it of the various phenols. As a result, it selected a total of nine phenol-type biophores. They are either simple thiophenols, as is the case with biophore **12**, or (thio)phenols substituted in (both) ortho or para positions with electron-withdrawing groups (NO_2, NO, Cl), as in biophores **5**, **6**, **13**, **18**, **19**, and **20**. Biophore **5** did not include the bromo derivatives because their distribution between actives and inactives was not skewed toward actives; therefore, the program instead selected biophore **13** as responsible for the acidity of 2-bromophenols.

The acidity of hinokiols is explained by biophore **3**, contained in 19 cyclic compounds. It contains the acidic enol group, but does not show the stabilization of the negative charge in the conjugate base by the conjugated double bonds in the seven-membered ring.

The program explained the acidity of heterocyclic compounds by selecting biophores **4** (for tetrazoles), **7** (for oxa- and thiazolidine-2,4-diones), **8** (for barbituric acids), and **9** (for xanthines). In all of them the negative charge (on nitrogen or carbon) in the conjugate base is stabilized by electron-withdrawing groups (N=N, CO), especially through conjugation.

Pseudoacids, with their acidic proton on carbon, are also identified by our program. Biophores **16**, **21**, and **22** stand for the acidity of tri- and disubstituted methanes. Tricyanomethane is separately explained by the last biophore, even though it might have been treated by biophore **16**. This is due to the way in which the bioisosteres are defined in the program. Barbituric acids, also pseudoacids, were discussed earlier.

One special note concerns biophore **11**, proposed to be responsible for the acidity of some polyfluoroalcohols. This biophore does not contain the acidic hydroxyl proton and results from a program misjudgment of the moiety responsible for the acidity of these compounds.

The modulators that affect the acidity of the carboxylic group in our database are given in Table III. They were selected upon performing a multivariate regression analysis (CASE analysis) on the subset of acids containing the carboxyl group with its expanded fragments. The QSAR equation given by the program is

$$pK_b = \text{constant} + \sum r_i(n_i M_i)$$

TABLE III.
List of Modulators Related to the COOH Biophore.

No.	Modulator	QSAR regression coefficient, r_i
0	constant	56.8
1	CS—SH	+16.4
2	NO_2—NH—	−5.5
3	SO_2—OH	+13.8
4	Cl—CH—	+9.1
5	Br—CH—	+9.3
6	SO_2—CH—	+6.8
7	CH_2—NH—CH—	+12.0
8	N=CH—CH=	−8.6
9	HO—CH_2—CH_2—	−12.5
10	HO—PO—C=	+5.9
11	CO—CO—OH	+6.6
12	CH_2—CH(CO)—CH_2—	−5.9
13	HO—PO(OH)—CH_2—	+3.4
14	CH_3—C—CH_2—CH_2—	−3.1
15	NH—CH_2—CH_2—NH—	−14.4
16	HO—CO—CH—NH_2	+7.3
17	F—C—CO—OH	+8.8
18	Cl—C—CO—OH	+8.3
19	HO—CO—CH_2—CH—CH_3	−4.9
20	HO—CO—CH_2—CH_2—C=	−5.8
21	HO—CO—CH_2—CH_2—CH=	−8.6
22	HO—CO—CH_2—CH_2—NH—	−5.1
23	Cl—C=C—CO—OH	+10.5
24	CO—CH_2—CH_2—CO—OH	−5.3
25	O_2N—C=C—CO—OH	+10.9
26	HO—CO—C(CO)—CH_2—CH_3	+3.5
27	HO—CO—C=C—CO—OH	+2.4
28	HO—CO—C=C—CH=CH—	+8.4
29	log *P*	−1.6
30	$(\varepsilon_{HOMO} - \varepsilon_{LUMO})/2$	−5.1

The order is that given by the program (increasing number of heavy atoms within each fragment).

where r_i represents the partial regression coefficient (see Table III), n_i is the frequency of occurrence of the modulating fragment M_i in a compound, and M_i indicates the presence or absence of the fragment i in the QSAR analysis.[§] For physicochemical or quantum mechanical descriptors, n_i is obviously 1 and M_i stands for the value of the property, which is also calculated by the program. For example, to predict the acidity of chlorodifluoroacetic acid (experimental pK_a is

[§] The experimental acidities ($pK_a = 14 - pK_b$) are translated by the program in internal activity units (Multi-CASE units, C) to fulfill the ranges imposed by the algorithm. Thus, for the 2464-acid training set, $pK_b = 5.0 + 198.0*[(C - 10)/(C + 1460)]$.

0.46), not included in the learning set, the program uses the preceding QSAR equation to calculate a pK_a value of 0.60 (see Scheme I). In the case of molecules containing several biophores, the program is able to predict the pK_a of each of them.

The most activating modulators (having a positive and relatively large value of their regression coefficient) reflect the acid-strengthening effect of electron-withdrawing groups (CS, PO, NO_2, SO_2, F, Cl, Br). Those with an opposite effect may be explained as stabilizing the COOH group by steric hindrance or hydrogen bonding. But some of them need further explanation. The contribution of the modulators that are expanded fragments of the COOH biophore (the first three fragments in Table III) is either positive or negative depending on the average activity of compounds that contain them. Aminoacids contain modulators that affect their acidities depending on whether they have Zwitterionic (acidic pK_a less than basic pK_a) or amphoteric character. As a result, modulators **7** and **16** increase the acidity of the COOH group, whereas **8** (encountered in the pyridine carboxylic acids) has a negative effect. Some amino-containing modulators (**15, 22**) have an unexpected negative contribution because they are the building blocks of long-chain molecules (such as ethylenediamino-diacetic or -dipropionic acid), where solubility and steric access of water toward the COOH group play an important role in determining the acidity. Modulator **27** is picked up by the program to explain the extra acidity of benzene-*o*-dicarboxylic acids, such as phthalic acid.

The logarithm of the octanol/water partition coefficient was identified by the program as being relevant, but its contribution is practically insignificant to acidity. However, it helps somehow in lowering the acidity of long-chain isomers, which have a lower aqueous solubility.

The quantum mechanical parameters—charge densities, the energies and the coefficients of the HOMO and LUMO orbitals—for each molecule in the database were calculated by using a simple HMO approach.[17] We chose a rather low-level method to deal with this large database for computational reasons. Instead of HOMO and LUMO energies, we used their semisum (absolute electronegativity) and semidifference (absolute hardness), respectively,[16] because they may be more appropriate for chemical interpretation. In this particular case, only one quantum mechanical descriptor was selected by the program to account for the acidity of carboxylic acids. The absolute hardness comes with a negative contribution (see Table III), suggesting that the acidity decreases with increasing hardness or decreasing softness. Indeed, the larger the difference between the energies of the frontier orbitals, the less likely for the acid to dissociate.

The statistics of the regression analysis for the carboxylic group, as well as for the other biophores, are given in Table IV.

The hierarchical scheme resulting from the Multi-CASE analysis of the acidity of the 2464 compounds was used to predict the acidity of 214 drugs whose pK_a's are available.[10] This database, tabulated by Newton and Kluza, contains a total of 237 acidic drugs, from which 23 molecules were taken out for the following reasons: Three contained elements other than those mentioned earlier (specifically, arsenic and mercury), six are charged molecules, six have structures too complicated to be dealt with by the program, two were not identified by name, two are duplicates, two were wrongly entered in the acid group, and two lacked data for the most acidic group. The composition of this database is diverse and includes cephalosporins, penicillins, sulfonamides, tetracyclines,

Modulator	r_i	n_i	M_i	$r_i(n_iM_i)$
F—C—CO—OH	+8.8	2	1	17.6
Cl—C—CO—OH	+8.3	1	1	8.3
log *P*	−1.6	1	1.05	−1.7
Hardness, $(\varepsilon_{HOMO} - \varepsilon_{LUMO})/2$	+5.1	1	1.14	−5.8
Total				18.4

Constant = 56.8;
pK_b = 56.8 + 18.4 = 75.2 Multi-CASE units = 13.4 pK units;
pK_a = 14.0 − pK_b = 0.6;
cf. experimental pK_a = 0.46;
error = 0.14 pK units.

SCHEME 1. The multi-CASE prediction of the acidity in the case of chlorodifluoroacetic acid.

TABLE IV.
The Statistics for the Local QSAR Equations Associated with the Acidic Biophores (the Biophore Structure is Given in Table II).

Biophore	*n*	*r*	SD	*m*
1	601	.800	0.55	30
2	57	.945	0.72	9
3	19	.973	0.39	5
4	5	.927	0.68	1
5	12	.943	0.49	2
6	8	.889	0.48	2
7	8	.997	0.07	3
8	4	1	0	1
9	3	1	0	1
10	2	—	—	—
11	2	—	—	—
12	6	.992	0.10	2
13	5	.950	0.78	1
14	1	—	—	—
15	1	—	—	—
16	4	.975	0.74	1
17	1	—	—	—
18	1	—	—	—
19	1	—	—	—
20	1	—	—	—
21	2	—	—	—
22	1	—	—	—

n is the number of compounds in the training set containing the given biophore, *r* is the regression coefficient, SD is the standard deviation, and *m* is the number of modulators per biophore.

butazones, derivatives of folic acid, benzodiazepins, and coumarin-type molecules. About half the database has no corespondent in the training set. Overall, only 22 molecules are common to both the training and test sets. The statistical parameters obtained by comparing calculated and experimental data for all 214 drugs are listed as Set 1 in Table V. As can be seen, the correlation coefficient *r* is .838, which, given the highly noncongeneric profile of the database, is acceptable. When the 22 common compounds were removed (see Table V, Set 2), the program gave almost the same results, with slightly decreased values of the statistical parameters. Most of the error is due to the large discrepancies found for a few compounds. An example is given by the butazones (4-monosubstituted pyrazolidine-3,5-diones), whose calculated pK_a values differ by more than three pK_a units from the experimental pK_a's (see Table VI). This is because the program was unable to identify the biophore responsible for their acidities and the compounds were predicted to be nonacidic. To prove the ability of the program to gain knowledge, we added to the training set 10 pyrazolidine-3,5-diones[20] and obtained an improved training set, which was again submitted to the Multi-CASE analysis (see the statistical results in Table V, Set 3). The newly updated dictionary of structural attributes contained one more biophore, which is the fragment CO—CH—CO, responsible for the acidity of butazones. Taking into account the fact that the training set can be continuously increased as the program learns additional information, we consider our method a convenient way to predict the acidity of new molecules.

Conclusion

We applied our Multi-CASE methodology to correlate the structure and the pK_a values of a large database of organic molecules. The database was obtained using available literature data on acidity. We did not include, for example, most of the compounds of the extensive compilation of pK_a values for acids published by Serjeant and Dempsey[20] due to limited computational capabilities at this time. In general, we find that the larger the database, the higher the predictive power of our program, so we intend to add more data to our training set (as computing resources increase) and expect an even better predictive power. However, even with the current 2464-molecule database, we were able to predict acceptably

TABLE V.
Statistical Parameters of the Multi-CASE Predictions of the Acidity of Drugs.

Set	Training / Test	Phi^2	OC, %	Sens, %	Spec, %	*r*	SD
1	2464 / 214	0.671	90.5	85.3	98.6	.838	1.52
2	2464 / 192	0.633	89.4	84.4	98.3	.824	1.58
3	2474 / 214	0.716	92.1	87.9	98.6	.858	1.44

TABLE VI.
The pK_a Values of Three Butazones Contained in the Test Set, before and after Adding 10 Pyrazolidinediones to the 2464-Acid Training Set.

Molecule	Exp. pK_a	Calc. pK_a		Error after training
		Before training	After training	
Oxyphenbutazone	4.7	9.0	4.14	−0.56
Phenylbutazone	4.5	9.0	4.14	−0.36
Sulfynpyrazone	2.8	9.0	4.14	+1.34

(r = .824) the acidity of 192 drugs, all of them structurally different from the compounds used in the training set. The program and its pK_a database are available from Professor Klopman.

Acknowledgments

This work was supported in part by the National Institute of Allergy and Infectious Diseases as a National Cooperative Drug Discovery Group for Opportunistic Infectious grant (AI 30189 SRC). The authors are indebted to Dr. Mario Dimayuga (Biofor Inc.) for help with computational problems.

References

1. T. H. Lowry and K. S. Richardson, *Mechanism and Theory in Organic Chemistry*, 3rd ed., Harper & Row, New York, 1987, p. 259.
2. D. D. Perrin, B. Dempsey, and E. P. Serjeant, pK_a *Prediction for Organic Acids and Bases*, Chapman and Hall, London, 1981; at the time we finished writing this article, Dixon and Jurs published their work about the estimation of pK_a for organic oxyacids using calculated atomic charges (S. L. Dixon and P. C. Jurs, *J. Comp. Chem.*, **14,** 1460 (1993)).
3. C. Grüber and V. Buss, *Chemosphere*, **19,** 1595 (1989).
4. K. Ohta, *Bull. Chem. Soc. Jpn.*, **65,** 2543 (1992).
5. W. L. Jorgensen and J. M. Briggs, *J. Am. Chem. Soc.*, **111,** 4190 (1989).
6. C. Lim, D. Bashford, and M. Karplus, *J. Phys. Chem.*, **95,** 5610 (1991).
7. G. Klopman, *Quant. Struct.-Act. Relat.*, **11,** 176 (1992).
8. G. Kortüm, W. Vogel, and K. Andrussow, *Dissociation Constants of Organic Acids in Aqueous Solution*, Butterworths, London, 1961 (G. Kortüm, W. Vogel, and K. Andrussow, *Pure Appl. Chem.*, **1,** 187 (1960)).
9. (a) J. March, *Advanced Organic Chemistry*, 4th ed., McGraw-Hill, New York, 1992, p. 249; (b) R. Stewart, *The proton: Applications to Organic Chemistry*, Academic Press, New York, 1985; (c) R. C. Weast, Ed., *CRC Handbook of Chemistry and Physics*, 60th ed., CRC Press, Boca Raton, FL, 1979, p. D161; (d) H. Mahler and E. Cordes, *Biological Chemistry*, 2nd ed., Harper & Row, New York, 1971; (e) A. Collumeau, *Bull. Soc. Chim. Fr.*, 5087 (1968); (f) H. C. Brown, D. H. McDaniel, and O. Haflinger, In *Determination of Organic Structures by Physical Methods*, Vol. 1, E. A. Braude and F. C. Nachod, Eds., Academic Press, New York, 1955, p. 567; (g) G. E. K. Branch and M. Calvin, *The Theory of Organic Chemistry*, Prentice-Hall, New York, 1941.
10. D. W. Newton and R. B. Kluza, In *Principles of Medicinal Chemistry*, 3rd ed., W. O. Foye, Ed., Lea & Febiger, Philadelphia, 1989, p. 861.
11. T. W. Schultz, *Bull. Environ. Contam. Toxicol.*, **38,** 994 (1987).
12. H. Kubinyi, In *Physical Property Prediction in Organic Chemistry*, C. Jochum, M. G. Hicks, and J. Sunkel, Eds., Springer-Verlag, Berlin, 1988, p. 235.
13. G. Klopman and S. Wang, *J. Comp. Chem.*, **12,** 1025 (1991).
14. G. Klopman and M. McGonigal, *J. Chem. Inf. Comp. Sci.*, **21,** 48 (1981).
15. G. Klopman, S. Wang, and D. M. Balthasar, *J. Chem. Inf. Comp. Sci.*, **32,** 474 (1992).
16. R. G. Pearson, *Proc. Natl. Acad. Sci. USA*, **83,** 8440 (1986).
17. F. A. Van-Catledge, *J. Org. Chem.*, **45,** 4801 (1980).
18. D. D. Perrin, *Dissociation Constants of Organic Bases in Aqueous Solution*, Butterworths, London, 1965.
19. J. R. McClintic, In *Comprehensive Medicinal Chemistry*, Vol. 1, C. Hansch, Ed., Pergamon, Oxford, 1990, p. 163.
20. E. P. Serjeant and B. Dempsey, *Ionization Constants of Organic Acids in Aqueous Solution*, Pergamon, Oxford, 1979.