The MultiCASE Program II. Baseline Activity Identification Algorithm (BAIA)†

Gilles Klopman

Chemistry Department, Case Western Reserve University, Cleveland, Ohio 44106

Received September 5, 1997⊗

A new algorithm is presented that allows one to uncover underlying relations between the biological activity of a diverse set of molecules and a global parameter such as the partition coefficient, solubility, and/or the redox properties.

INTRODUCTION

When evaluating the relationship that may exist between the structures of a set of diverse molecules and their biological activity, one is sometimes confronted with the observation that some molecules are active in spite of the fact that they do not possess any well-defined functionality that may be related to the activity. This activity may be due to some global molecular property responsible for marginal activity, such as transport properties, mass or size properties, oxido-reducing properties, and so on. An example of such phenomenon is that of toxicity to fish,[1,2] where it was postulated that in addition to the specific toxicity that some chemicals may exert on fish, the mere lipophilic properties of chemicals may be sufficient to explain a "baseline" toxicity due to the so-called narcotic effect of molecules whereby lipophilic molecules induce sleep and the fish drowns.

This lipophilic effect is easily handled in congeneric databases where a partition effect parameter can be used to help explain observed activities. The number of "Hansch" type correlations of that kind[3,4] is testimony to the importance of this approach. In the case of diverse molecules, this baseline activity may be hidden by the mass of molecules that may possess certain functionalities giving them specific toxic (or pharmacological) activity.

Our group has developed over the years a "knowledge based program" called CASE[5] (Computer Automated Structure Evaluation) and, in its last version, MultiCASE[6] (MCASE) that basically transforms the knowledge of structure and activity of large sets of molecules into expert systems capable of predicting the activity of molecules not tested as yet. Both CASE and MCASE have the capability of calculating and using the water/octanol partition coefficient log P[7] if necessary to improve internally calculated quantitative structure−activity relationships (QSAR). CASE does this within the context of a global equation that attempts to correlate the entire database with the structure of the constituting molecules. MCASE, on the other hand, first reclassifies the molecules into smaller subsets, based on the observed presence of automatically identified biophores (pharmacophores or toxicophores), and uses log P values if necessary as modulators of the activity of these automatically selected congeneric subsets of molecules.

Neither of these procedures however provides a satisfactory solution to the problem. CASE often fails to recognize the importance of log P because it is buried within many other descriptors. MCASE does not recognize the problem at all because, if a molecule does not contain a biophore, its activity is presumed to be zero.

In this paper, we describe a simple baseline activity identification algorithm (BAIA) capable of uncovering a hidden correlation with a global parameters such as log P in highly diverse databases.

METHODOLOGY

The problem as it is normally seen is that a large number of diverse molecules and varying biological activities do not seem to follow any linear or other relationship with a property that is calculated for these molecules. The reasons for this can be diverse; for example, there may not be a relationship or there are other reasons why the molecules show activity or the activity produced by the relationship with the calculated property is weak compared with the activity produced by other variables.

The rationale we propose to follow is simple and based on the fact that if indeed there is a baseline activity, all the activity deviations from that baseline would be positive. In other terms, molecules would nearly always be more active than predicted by the baseline relationship. Further, the largest deviations from the baseline would be due to some factors other than baseline activity. Typically, positive deviations would be produced by specific functionalities responsible for increased activity. Therefore, the algorithm is as follows.

A linear regression is sought between observed activity and the calculated property that may produce a baseline. The calculated activity values, calculated from the regression analysis, are compared with the actual activity values. The molecule showing the largest deviation is eliminated and a new regression analysis is performed. The process is repeated until the F-test of the correlation cannot be improved by the dismissal of the next molecule or until one third of the molecules of the database have been eliminated. The procedure is then stopped and evaluated.

For the correlation to be accepted, it must satisfy a number of requirements. First, it must show sufficient statistical

† CASE and MultiCASE (MCASE) are trademarks of MULTICASE Inc., PO 22517, Beachwood, OH 44122.

⊗ Abstract published in *Advance ACS Abstracts*, December 15, 1997.

S0095-2338(97)00079-6 CCC: $15.00
Published on Web 01/19/1998

significance as measured by the *F*-test. Secondly, the slope of the correlation should be significant. Indeed, if the correlation is a horizontal line (i.e., slope $\cong 0$), then clearly the relationship is of no significance. There is one complication, however, which is due to the fact that the calculated property (e.g., the log of the water/octanol partition coefficient) may in some cases be substantially incorrect. This incorrectness may result in a positive or negative deviation about the calculated value of the activity for that molecule. If the deviation is positive, then the molecule will appear as more reactive than anticipated and the remaining treatment of the data will have to deal with it as if it contained a biophore. However, if the deviation is negative, it will erroneously detract from the existence of a baseline regression. For this reason, the algorithm is allowed to eliminate molecules that are less active than predicted by the baseline correlation with a warning that the calculated property may be incorrect. Thus, the third requirement is that although less active molecules can also be eliminated from the correlation, their number should not exceed a certain limit, which we choose arbitrarily to be 10% of the total number of molecules eliminated from the correlation.

Once the baseline has been identified, the activity of all the molecules is recalculated as the activity above baseline, and the normal MCASE procedure is then used to evaluate the relationship between the structure of the molecules and their excess activity.

In the following section, we describe the application of our models to four databases, two where the new algorithm kicks in, and the other two, where no global relationship with log *P* is detected. These four databases had been studied previously, sometimes with fewer molecules, by the CASE or MCASE methodologies. The four databases are: (1) *In vitro* inhibition of sparteine monooxygenase[8]; (2) in vivo eye irritation[9]; (3) mutagenicity of chemicals in Salmonella *typhimurium*[10]; and (4) agonist activity of capsaicin analogues.[11] The objective of the exercise was to determine whether a baseline correlation existed with the log of the octanol/water partition coefficient, log *P*.

APPLICATIONS

A. In Vitro Inhibition of Sparteine Monooxygenase. We have previously used the MCASE program to evaluate the structural factors responsible for the activity of compounds found to inhibit the activity of sparteine monooxygenase.[8] We refer the readers to the resulting paper[8] where the 74 compounds used for the study are listed and where a description of the activity of sparteine and the importance of preventing its oxidation are described. That study identified four fragments believed to be linked to activity. Three of them, when present in molecules, activate their ability to inhibit oxidation, whereas one fragment, when present, prevents activity. A major finding of that study was the importance of the water/octanol partition coefficient. Indeed, a high value of that coefficient was apparently sufficient to convey inhibitory properties to the molecules that possess it.

We selected that database for this study to see if our new BAIA will discover this relationship and to find how this affects the remaining part of the program, namely the identification of functionalities (biophores) relevant to activity. For consistency, we kept the same nomenclature and numbering of molecules as in the original paper.

As in the previous study, we used the same 58 compounds as our training set and kept the randomly selected 16 others for validation purposes. Activities were given in "CASE" units, where inactive molecules were given an activity of 15, marginally active molecules a value of 25, and active, very active, and extremely active molecules were given values of 35, 45, and 55, respectively. Admittedly, this ranking is not very quantitative and should present a challenge to any attempt to rationalize the data quantitatively. Nevertheless, we submitted the 58 compounds to analysis.

When the database was submitted to the MCASE program outfitted with the BAIA, the program immediately identified some relevance of the log *P* value by uncovering a global relation between the molecules of the database and the program-calculated log *P* values. The *F*-test however was only 19.3. The program then proceeded to refine the correlation by eliminating those molecules showing the greatest deviation from the value obtained from the linear fit. In all, 20 molecules were eliminated before the programming criteria were satisfied. At that point 38 molecules remained in the dataset. For these 38 molecules, the correlation coefficient *R* was 0.857 and the *F*-test climbed to 99.3. The standard deviation for the correlation was 6.17 (activities ranging from 15 to 55), and the linear regression equation was as follows:

$$\text{activity} = 11.42 + 6.49 \log P \qquad (1)$$

The relationship is both significant and important. Indeed, the *F*-test value is substantial and the coefficient of log *P* in the equation is large, indicating a strong relation with lipophilicity. Thus, it only takes a 1.5 change in the log *P* value to move a molecule from one activity category to another.

An analysis of the events that took place shows that of the 58 molecules that were evaluated, 13 were more active than predicted by the linear fit and seven were less active than predicted. This result fits our pre-analysis criteria that assumed that those molecules that are more active than predicted probably possess some other properties that makes them particularly active, and those molecules that are less active possess some properties that prevent them from exhibiting activity. Either of these classes of molecules could alternatively end up in such a category if the partition coefficients are grossly miscalculated. The fact that the ratio of molecules more active than calculated from baseline over those that are less active than baseline is $\sim$2:1 is convincing evidence that in addition to baseline activity, specific activity exists that, if identified, could be used to create superior inhibitors.

After accepting the conclusions of the BAIA, MCASE proceeded to identify the fragments (called biophores) that may be responsible for such specific activity and identified four fragments as being responsible for the activity of the 20 molecules that have substantial activity above the baseline of activity due to lipophilicity.

We then proceeded to predict the activity of the 16 compounds that were withdrawn from the learning set. The results are shown in Table 1.

Table 1. Prediction of the Inhibitory Activity of 16 Compounds Not Included in the Learning Set

no.	molecule	exp	calc[a]	calc[b]
1	amitriptyline	++	+	++
2	imipramine	++	++	+++
3	phenytoin	−	+	+
4	trimethadione	+	(−)	(−)
5	metropolol	++	+	+
6	labetalol	+++	(++)	(+)
7	timolol	+++	(−)	(+)
8	cinchonine	+++	(++)	(++)
9	quinine	+++	++	+++
10	pimozide	+	+++	+++
11	chloroquine	++	+++	+++
12	17-*n*-pentylsparteine	+++	+++	+++
13	debrisoquine	++	(+)	(+)
14	hexamethonium	−	(+++)	(+++)
15	guanoxan	+++	(+++)	(+++)
16	quinacrine	++	+++	+++

[a] Calculated by MCASE using internally calculated log *P* values (see Table 2). [b] Calculated by MCASE using experimental log *P* values (see Table 2); (xx) indicates that the molecule contains a certain structural feature not encountered in the data base and whose importance can therefore not be assessed (the predictions are therefore somewhat less reliable).

Table 2. Calculated and Experimental Values of log *P* of the Tested 16 Compounds Not Included in the Learning Set

no.	molecule	log *P* calc	log *P* experimental
1	amitriptyline	2.59	3.94
2	imipramine	3.86	4.62
3	phenytoin	2.50	2.40
4	trimethadione	−0.43	−0.37
5	metropolol	2.21	2.34
6	labetalol	3.32	2.51
7	timolol	0.85	1.91
8	cinchonine	3.29	−
9	quinine	3.48	4.83
10	pimozide	6.21	6.30
11	chloroquine	5.60	4.63
12	17-*n*-pentylsparteine	4.08	−
13	debrisoquine	2.71	−
14	hexamethonium	4.81	−
15	guanoxan	4.47	−
16	quinacrine	6.69	−

As can be seen, nine of the 12 active or very active molecules are predicted correctly and two out of the four inactive or marginally active molecules are also recognized. It is to be noted, however, that the program qualified its predictions in 11 of the 16 compounds by recognizing that some features of the molecules that may be relevant to activity were never documented by the learning set or that its predictions are based on statistically poorly documented features. Labetalol was particularly peculiar because quantitative structure−activity relationship (QSAR) analysis predicted it to be inactive, but the program also indicated that it had 82% chance of being active.

Overall, the results are comparable with those obtained in the original work. However, in the original work, experimental values of log *P* were used, whereas here, all values are calculated by the program. It turns out that some of these values differ significantly from experimental values (Table 2).

In Vivo Eye Irritation.[9] This database consisted of the 186 compounds of our previous study, augmented with 21 compounds whose eye toxicity data became available after the original study was completed. As before, the eye toxicity of the compounds was ranked by giving an index of 15 to nontoxic molecules, 25 to those listed as marginally toxic, 35 to the toxic molecules, 45 to the molecules that were listed as very toxic, and 55 to those found to be extremely toxic. As already mentioned, it is clear that this ranking is not very quantitative.

The initial correlation between the activities and the log *P* of the molecules gave an *F*-test of 28.9. This value is surprisingly good, considering the qualitative nature of the activity data. Upon elimination of 40 molecules, of which 30 molecules were more active than calculated from the best correlation with log *P* and 10 molecules were less active than calculated, we obtained a final correlation between log *P* and the remaining 167 molecules with a correlation coefficient *R* equal to 0.60 and an *F*-test of 93.2. The standard deviation of the residuals was 13.637 (activities ranking from 15 to 55) and the equation was as follows:

$$\text{activity} = 34.5 - 3.155 \log P \qquad (2)$$

This equation indicates that molecules tend to be more toxic when they are hydrophilic. However, the change in activity with changes in log *P* are rather small (i.e., one notch in the activity scale for a change of 3 units of log *P*). Overall, one may conclude that the relationship is highly relevant but of relatively little consequence. The relationship may simply indicate that for the toxicity of a compound to be expressed, it must be relatively soluble in the water phase.

Once the relationship with log *P* has been established, the program proceeds to identify the biophores needed to explain the toxic activity of the molecules that are significantly more active than predicted by the log *P* correlation. In this case, the program identified 20 biophores compared with the 37 that were needed to explain the slightly smaller database using MCASE without BAIA.

Mutagenicity of Chemicals in Salmonella *typhimurium*.[10] A database of 1353 compounds, tested for mutagenicity in Salmonella *typhimurium* by the National Toxicology Project (NTP) was submitted for further analysis. This database is an outgrow of the database that we studied with the CASE program in 1990.

The activities are simply given as active or inactive, and therefore there is not much hope to obtain significant correlations beyond the expert system type of predictions ordinarily given by the CASE and MCASE programs. Nevertheless, we submitted the database to analysis and found an initially poor correlation with log *P*. However, as active molecules are removed, the correlation progressively improves, reaching an *F*-test value of 137 when 332 active and one inactive molecules have been removed. The correlation coefficient, to no surprise, remained low (i.e., 0.345) and the standard deviation was 9.75 ("CASE" activity values of 10 were initially assigned to inactive molecules and the mutagenic ones received a score of 39). The correlation equation was as follows:

$$\text{activity} = 17.91 - 1.44 \log P \qquad (3)$$

As can be seen from these numbers, the correlation is highly significant but of rather small importance. Indeed, it would take a log *P* difference of 10 units to move a molecule from the active to the inactive list or vice versa. Nevertheless, a

correlation does exist showing that hydrophilicity is favorable to the observation of mutation in Salmonella.

This correlation was refused by our algorithm because of the low (absolute) value of the coefficient of log *P*. When the analysis of the data proceeded further to identify the biophores responsible for activity, the results were about the same, whether the correlation with log *P* is or is not accepted.

Agonist Activity of Capsaicin Analogues.[11] The quantitative agonist activity of 123 capsaicin analogues was studied in our laboratory in 1995.[11] The analysis resulted in the identification of three highly relevant biophores. Taken together, they accounted for the agonist activity of 59 active molecules of the 66 active molecules of the database. The activities here are expressed as concentration that is effective in 50% of the cases (EC_{50}) in μmole of maximal Ca^{+2} influx into neonatal rat dorsal root ganglia. It was considered that molecules are basically inactive if their EC_{50} was greater than 100 μmol. Thus, of the 123 compounds, 66 were classified as actives and 57 as inactives.

When the database was submitted to the program, the *F*-test for correlation with log *P* was only 4.5. That number did not increase when the program attempted to remove outliers and the program therefore bypassed any possible correlation with the partition coefficient. The CASE analysis continued normally and produced the same result as described in our previous communication on this subject.

CONCLUSION

We have shown that the BAIA is capable of recognizing the existence of a baseline relationship between activity and the log *P* of a set of diverse molecules. The procedure is automatic and reproducible. In followup papers, we will show that this technique can be used to study the toxic effect of diverse chemicals to marine life as well as other toxic and pharmacological endpoints.

REFERENCES AND NOTES

(1) Van Leeuwen, C. J.; Van der Zandt, P. T.; Aldenberg, T.; Verhaar, H. J.; Hermens, J. L. *Sci. Total Environ.* **1991**, *109*−*110*, 681−690.
(2) Vittozzi, L.; De Angelis, G. A. *Aquatic Toxicol.* **1991**, *19*, 167−204.
(3) Hansch, C.; Hoekman, D.; Gao, H. *Chem. Rev.* **1996**, *96*, 1045−1075.
(4) Hansch, C.; A. *Substituent constants for correlation analysis in chemistry and biology.*; John Wiley and Sons: New York, 1979.
(5) Klopman, G. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7320.
(6) Klopman, G. *Quanti. Struct.-Act. Relat.* **1992**, *11*, 176−184.
(7) Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752−781.
(8) Klopman, G.; Venegas, R. E. *Acta Pharma. Jugosl.* **1986**, *36*, 189−208.
(9) Klopman, G.; Ptchelintsev, D.; Frierson, M. R.; Pennisi, S.; Renskers, K.; Dickens, M. *ATLA* **1993**, *21*, 14−27.
(10) Rosenkranz, H. S.; Klopman, G. *Mutation Res.* **1990**, *228*, 51−80.
(11) Klopman, G.; Li, J.-Y. *J. Computer-Aided Molecular Design* **1995**, *9*, 283−294.

CI9700790