Journal of Computer-Aided Molecular Design **18:** 135–143, 2004.

Quantitative structure–activity relationships from optimised *ab initio* bond lengths: steroid binding affinity and antibacterial activity of nitrofuran derivatives

P.J. Smith & P.L.A. Popelier*
Department of Chemistry, UMIST, Manchester M60 1QD, UK

Received 17 September 2003; accepted in revised form 24 March 2004

Key words: ab initio, active center, bond length, corticosteroids, nitrofurans, QSAR

Summary

The present day abundance of cheap computing power enables the use of quantum chemical *ab initio* data in Quantitative Structure–Activity Relationships (QSARs). Optimised bond lengths are a new such class of descriptors, which we have successfully used previously in representing electronic effects in medicinal and ecological QSARs (enzyme inhibitory activity, hydrolysis rate constants and pK_as). Here we use AM1 and HF/3-21G* bond lengths in conjunction with Partial Least Squares (PLS) and a Genetic Algorithm (GA) to predict the Corticosteroid-Binding Globulin (CBG) binding activity of the classic steroid data set, and the antibacterial activity of nitrofuran derivatives. The current procedure, which does not require molecular alignment, produces good r^2 and q^2 values. Moreover, it highlights regions in the common steroid skeleton deemed relevant to the active regions of the steroids and nitrofuran derivatives.

Introduction

The establishment of numerical relationships between biological activity and molecular structure, under the heading 'quantitative structure–activity relationships' (QSAR) [1], plays an important role in drug design. Many different approaches have been proposed to QSAR [2–4], each using various molecular descriptors in combination with different chemometric techniques. Traditional approaches to QSAR, pioneered by Hansch and Fujita [5], involve correlating two-dimensional empirical descriptors with activity. More recently, approaches involving three-dimensional structural data [6–10] have provided a powerful alternative to these extrathermodynamic parameters. Furthermore, relative similarity measures [11–13] between two molecules have enjoyed great success as a source of additional structural descriptors [14].

In recent years we have been interested in injecting data obtained from quantum mechanical calculations into QSARs, particularly quantum topological properties obtained from the theory of 'Atoms in Molecules' (AIM) [15–18]. We developed a technique called Quantum Topological Molecular Similarity (QTMS), which has delivered excellent QSARs, such as σ_p, σ_m, σ_I and σ_p^0 parameters of mono- [19] and poly-substituted benzoic acids, phenylacetic acids, bicyclo carboxylic acids [20]; toxicity and biodegradability of *p*-substituted phenols and ^{13}C NMR chemical shifts in *p*- and *m*-substituted benzonitriles [21]; antibacterial activity of nitrofuran derivatives; anti-tumor activity of (E)-1-phenyl-but-3-en-ones [22]; toxicity of polychlorinated dibenzo-p-dioxins (PCDDs) [20]; prediction of hydrolysis rate constants of polar esters [23]; estimation of pK_a of carboxylic acids, anilines and phenols [24]; and many others.

Early on in the development of QTMS we discovered [25] that the bond lengths of energy-optimised molecules (i.e. local minima or all positive eigenvalue stationary points on the potential energy surface) are first-rate descriptors, able to capture the electronic effects in a QSAR. In this study we do not invoke quantum topological properties but describe

*To whom correspondence should be addressed. Fax: +44-161-2004559; E-mail: pla@umist.ac.uk

a much simpler approach, involving only optimised bond lengths obtained from semi-empirical and *ab initio* calculations. The dramatic enhancement of computational power coupled with dwindling costs makes such an approach feasible. Our objective is to show that bond lengths alone can be used in producing good, valid QSAR models and are also capable of suggesting regions of the molecule responsible for a given activity.

In order to demonstrate the basic ideas of the approach, we have chosen to apply our method to the well-known steroid dataset [10]. This consists of a series of 31 steroids that bind to the corticosteroid-binding globulin (CBG) receptor. This same set of data has been widely investigated by different authors [8,26–30] and has become a benchmark against which the success of a QSAR can be gauged [31]. Many different techniques have been applied to this dataset in the past, with acceptable results. In this paper the dataset is used to illustrate that satisfactory correlations between biological activity and bond lengths can be achieved, in addition to highlighting the region of the molecule responsible for the activity.

Computational method

Initially an estimated geometry for each steroid is obtained from the program MOLDEN [32], and subsequently optimised using the *ab initio* program GAUSSIAN98 [33]. This program then performs a so-called single point calculation on the optimised geometry, which yields the electronic energy in addition to the wave function. Geometry optimisations and single point calculations were first carried out at the semi-empirical AM1 [34] level, using the default convergence criteria. The calculations were then repeated at a modest Hartree–Fock level, denoted by HF/3-21G* [35], in order to see if the bond lengths obtained from the *ab initio* method produced a significantly different and potentially superior model. Although the 3-21G* basis set is quite minimal, it is known to produce reliable geometries. By selecting a system of molecules with a rigid backbone, the complications arising from conformational flexibility are effectively eliminated. Obtaining the optimised geometries represents the most time consuming part of the QSAR's construction. On a moderately priced PC (dual AMD Athlon MP1900+, 1 GB DDR RAM) it took approximately 130 hours of CPU time to obtain all 31 steroid geometries at Hartree–Fock level, whilst the AM1 calculations were more than two orders of magnitude faster, requiring about half an hour of CPU time. This means that on a Linux PC cluster of merely six PCs all descriptors would be obtained in less than one full day. Chemometric analysis required less than 15 minutes of computing time.

Once the wave function files have been produced they are then read by (a local version of) the program MORPHY98 [36], which extracts the required bond lengths and exports them into a format that is convenient for subsequent statistical analysis. Using this program greatly facilitates the data organisation and is the primary reason wave functions are the method of choice for obtaining the molecular descriptors. Note that for a full QTMS analysis, including AIM's quantum topological descriptors, one needs the complete wave functions.

In a final stage a model is constructed using an advanced multi-linear regression technique known as Partial Least Squares (PLS) [37, 38] in conjunction with a genetic algorithm-based [39] variable selector. PLS generates its own special Principal Components, called Latent Variables (LV), which are linear combinations of the independent variables. The PLS analysis is performed by the program SIMCA-P [40], which yields four statistics that are used to judge both the quality and validity of the model. The first of these is the correlation coefficient, r^2, which assesses the 'goodness of fit' of the model. This statistic is potentially misleading, because it is possible to produce a good fit simply by employing a large number of descriptor variables. Hence, r^2 is used in conjunction with the cross-validated r^2, denoted by q^2. The latter coefficient is dependent on the PRESS score [41], calculated *by leaving out one seventh of the data*, and provides a measure of the predictive power of the model. It could be argued that this fairly substantial (4 compounds out of 31) and systematic omission of data (up to $31 \times 30 \times 29 \times 28/4! = 31{,}465$ possible subsets) offers a viable alternative to the often followed procedure of splitting the data set into a training set and a test set. The latter procedure has the merit of providing a stringent test to the predictive capacity of the training set, since the test set is usually quite large (sometimes a third of the size of the training set). On the other hand, the decision as to which compounds belong to the training or to the test set is usually quite arbitrary, an issue which is avoided by our current leave-group-out [30] style scheme. The final two statistics, obtained via a randomisation validation test, provide a safeguard against the possibility

of obtaining a model by chance. They are denoted by r^2_{int} and q^2_{int} where the subscript denotes 'intercept'. In the randomisation test the experimental data ('Y') are randomly permuted, followed by a PLS re-run. Each randomisation and subsequent PLS analysis generates a new r^2 and q^2 value, which is plotted against the absolute value of the correlation coefficient between the original set of dependent variables and its permutation. After 10 permutations a line is fitted through the r^2 values and another through the q^2 values, yielding two separate intercept values. A model is considered valid [37] if $r^2_{int} < 0.4$ and $q^2_{int} < 0.05$.

In order to interpret the model obtained it is appropriate to examine the Variables Important to the Projection (VIPs) [42] produced by SIMCA-P. The VIPs give the relative importance [43] of each descriptor contributing to the model. Bonds with higher VIP scores are considered more relevant in explaining the activity. Since each VIP is associated with a particular bond it is possible to recover the regions of the molecule responsible for causing the change in a given activity. Finally, the information obtained from the VIPs is projected onto a diagram of the common molecular skeleton of the dataset, via a colour code, in order to obtain a visual interpretation of the model. Variables with a VIP value of less than one are considered unimportant to the model [37] and can be discarded. The remaining variables are then assigned to one of five linearly spaced intervals, each associated with a colour. Bonds occurring in the interval representing the highest VIP values are assigned the colour red. Subsequent intervals of decreasing VIP values are given the colours yellow, green, blue, and magenta, respectively. The colour code indicates the *relative* importance of each bond to the model and expresses how well localised or diffuse the active region is. It is important to note that the colours do not express absolute values of VIPs and therefore cannot be used to compare plots between different QSAR models.

Once the initial PLS model incorporating all descriptors is obtained, variable selection is performed using a Genetic Algorithm (GA) in order to select the optimum number of descriptors for use in subsequent PLS analysis. A new PLS model is then obtained using the descriptors selected by the GA. Variable selection is not employed in the original analysis in order to compare the bonds selected by the GA with the bonds that are allocated high VIP scores in the model obtained using all the descriptors. For the GA application described here the MATLAB routine *genalg.m* from the PLS Toolbox [44] was employed. The GA involved a population of 256 randomly selected models, the mutation rate was set at 0.03, the maximum number of generations was set at 200, and the fitness function was the cross-validation error from PLS performed on the dataset using variables (bond lengths) selected by the GA.

Results and discussion

CBG binding affinity of corticosteroids

The regulation of human physiology depends crucially on corticosteroids [30], given their role in carbohydrate, lipid and protein metabolism. They also play an important role in pregnancy, scar tissue, Crohn's disease, the maintenance of proper electrolyte balance and as antirheumatic and antiallergenic agents.

The origin and history of the 'standard steroid' set that Cramer et al. [10] used to launch their CoMFA analysis is complicated, not in the least because of published multiple errors in topological coding and stereochemistry, as spotted and corrected by Wagener, Sadowski and Gasteiger [26]. We base the data in Table 1, containing the entire set of 31 steroids, on the clear and reliable exposition of Coats [31]. The CBG binding data are expressed as affinity constants K, which are conveniently converted to $\mathrm{pK} = -\log \mathrm{K}$ values. The more negative the pK value the higher the binding affinity. For example, cortisol, having a pK value of -7.881 shows one of the highest binding affinities in the set. Note that seven steroids have a CBG affinity constant of less than 0.1×10^6 M^{-1} corresponding to a pK value of higher than -5 (e.g. -4.7). Nevertheless, we adopt the minimum value of -5 in our analysis. The trivial names for all steroids have been adopted from Coats [31]. Because the steroid structures in this dataset have been presented as charts in numerous papers, they will not be included here. Although it is not necessary to perform any molecular alignments, an *a priori* numerical labeling of the atoms common to all the molecules is required. For convenience we adopt the IUPAC numbering convention for steroids. The common molecular skeleton is given in Figure 1. Note that the numerical label of O_{18} is not compatible with the IUPAC convention, but is introduced here as a convenient way of naming the 18th atom of the common skeleton.

The results of the regressions for both sets of bond lengths are shown in Table 2, whilst the plot of the observed versus predicted corticosteroid binding affinities for both models is given in Figure 2.

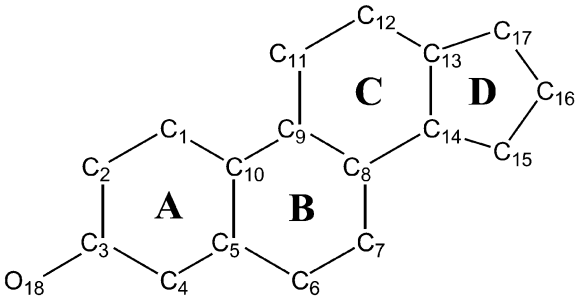

Figure 1. Numbering scheme for the steroid dataset.

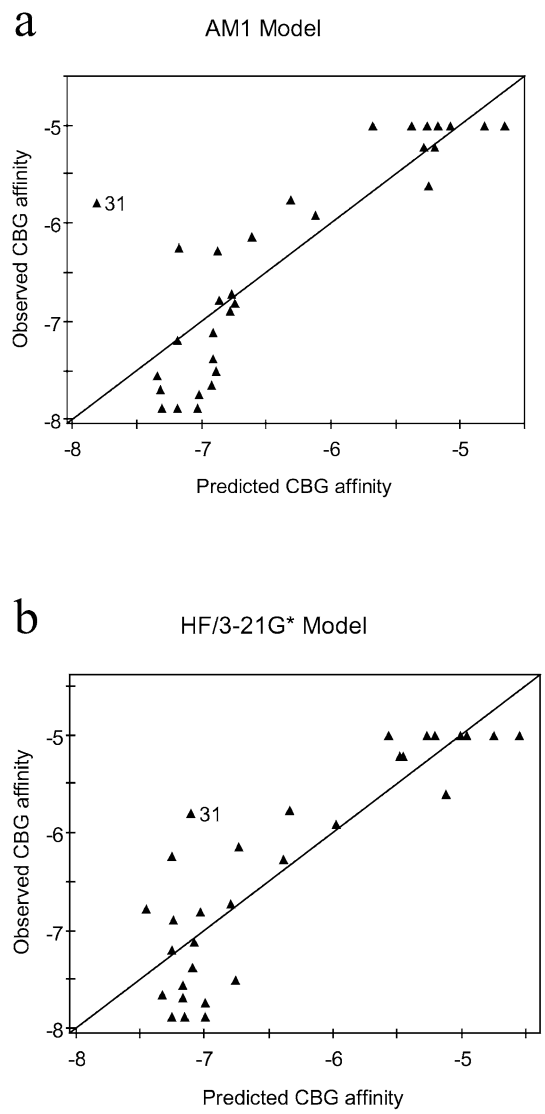

Figure 2. Observed versus predicted CBG binding affinities for the steroid set using optimised bond lengths as descriptors, obtained at (a) AM1 level, and (b) HF/3-21G* level.

Table 1. Steroid set: CBG binding affinity data.

No.	Compound	CBG ($K\times10^6$)	CBG (pK)
1	aldosterone	1.9	−6.279
2	androstanediol	<0.1	−5
3	androstenediol	<0.1	−5
4	androstenedione	0.58	−5.763
5	androsterone	0.41	−5.613
6	corticosterone	76	−7.881
7	cortisol	76	−7.881
8	cortisone	7.8	−6.892
9	dehydroepiandrosterone	<0.1	−5
10	deoxycorticosterone	45	−7.653
11	deoxycortisol	76	−7.881
12	dihydrotestosterone	0.83	−5.919
13	estradiol	<0.1	−5
14	estriol	<0.1	−5
15	estrone	<0.1	−5
16	etiocholanolone	0.18	−5.225
17	pregnenolone	0.18	−5.225
18	17-hydroxypregnenolone	<0.1	−5
19	progesterone	24	−7.380
20	17-hydroxyprogesterone	55	−7.740
21	testosterone	5.3	−6.724
22	prednisolone	–[a]	−7.512
23	cortisol-21-acetate	–	−7.553
24	4-pregnene-3,11,20-trione	–	−6.779
25	epicorticosterone	–	−7.200
26	19-nortestosterone	–	−6.144
27	16α,17-dihydroxyprogesterone	–	−6.247
28	16α-methylprogesterone	–	−7.120
29	19-norprogesterone	–	−6.817
30	2α-methylcortisol	–	−7.688
31	2α-methyl-9α-fluorocortisol	–	−5.797

[a]No measured values. pK taken from Table IVB in Ref. [10].

Table 2. PLS statistics using all descriptors.

Level	r^2	q^2	r^2_{int}	q^2_{int}
AM1	0.708	0.575	0.16	−0.03
HF/3-21G*	0.756	0.697	0.16	−0.04

It can be seen that both sets of bond lengths produce good, predictive models (with only 1 Latent Variable) that pass the validation criteria set by the randomisation test. Although the HF/3-21G* *ab initio* model produces a superior fit, it is gratifying to note that the semi-empirical bond lengths, produced at a fraction of the cost, still yield reasonable regression statistics. We emphasise that in the initial model we made no attempt to maximise any of the correlation

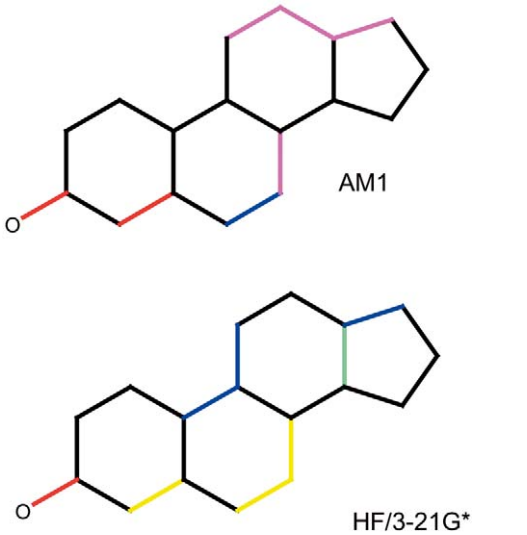

Figure 3. Colour-coded plot expressing the influence (via the VIP values; see text) of the bonds in explaining the observed CBG binding affinity of the steroids. For AM1 the highest VIP value (2.1) is found for the C_3O_{18} bond and the next highest is 1.8 (for the C_4C_5 bond), followed by a substantial drop to 1.3 for C_6C_7. The VIP profile for HF/3-21G* is smoother, starting with 1.8 as the highest VIP value.

coefficients by removing molecular outliers (such as the notorious steroid 31) or by selection of the original descriptor variables. The goal is to straightforwardly demonstrate that good QSAR models can be obtained using equilibrium bond lengths. However, if outlier 31 is omitted from the set then the r^2 and q^2 values become 0.86 and 0.81, respectively.

The colour-coded molecular skeletons in Figure 3, constructed using the VIP scores of each model, highlight the relative importance of each bond to the regression. Both models are in broad agreement. The diagrams clearly suggest that the biological activity of the steroids is related to the nature of the bond between atoms C_4 and C_5, and the carbon-oxygen bond occurring at atom C_3. Interestingly, examination of the steroid structures reveals that the most active compounds all possess a double bond between C_4 and C_5, in addition to a carbonyl group at C_3. Other bonds implicated by the colour plots have VIP scores that are lower than those found on the *A* and *B* rings and are unlikely to be as important in explaining the activity. The regions of the molecule suggested by the models overlap strongly with those suggested by Mickelson et al. [45], who conclude that a 3-oxo group is important for optimal binding. Reduction of this group to either a 3α-hydroxyl or 3β-hydroxyl results in significantly lower binding. The authors also state that reduction

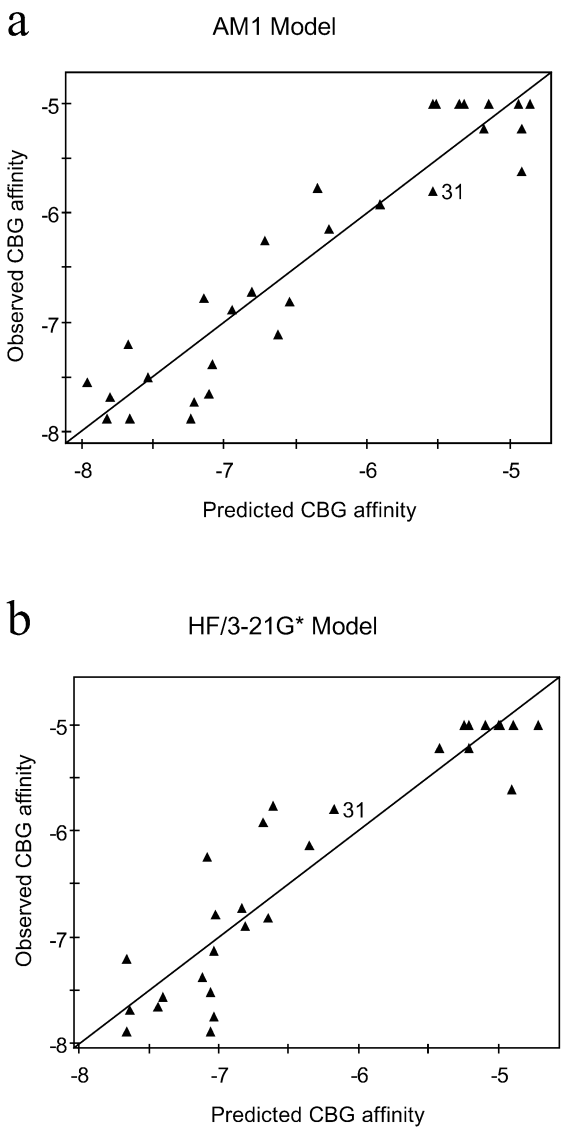

Figure 4. Observed versus predicted CBG binding affinities for the steroid set using optimised bond lengths, obtained at (a) AM1 level and (b) HF/3-21G* level, *after employing GA* variable selection.

of the double bond at the C_4 position to either the 5α or 5β derivative similarly produces a reduction in binding affinity. Similar conclusions about the active regions of the steroid molecules have been drawn from research involving E-state descriptors [30], quantum similarity measures [46], and hydrogen electrotopological state fields [47]. In a SOMFA study Robinson et al. [28] highlight a large area of negative potential, occurring at atom C_3, as being important in explaining the activity. It has also been observed that good correlations with CBG binding can be obtained with the

Table 3. PLS statistics from SIMCA-P using descriptors after GA variable selection.

Level	r^2	q^2	r^2_{int}	q^2_{int}
AM1	0.885	0.678	0.11	−0.11
HF/3-21G*	0.861	0.816	0.03	−0.14

Table 4. Nitrofuran derivative set: experimental 50% growth inhibitory activity values.

Compound	R	$\log(1/IC_{50})$	$\log(1/IC_{50})$
		S. aureus	*C. crescentus*
1	OCH_3	3.82	2.98
2	H	3.52	2.89
3	CH_3	3.94	2.90
4	C_2H_5	3.35	2.56
5	n-C_3H_7	3.35	2.51
6	Cl	3.24	2.28
7	Br	3.18	2.50
8	CN	2.50	2.34
9	NO_2	2.40	1.93

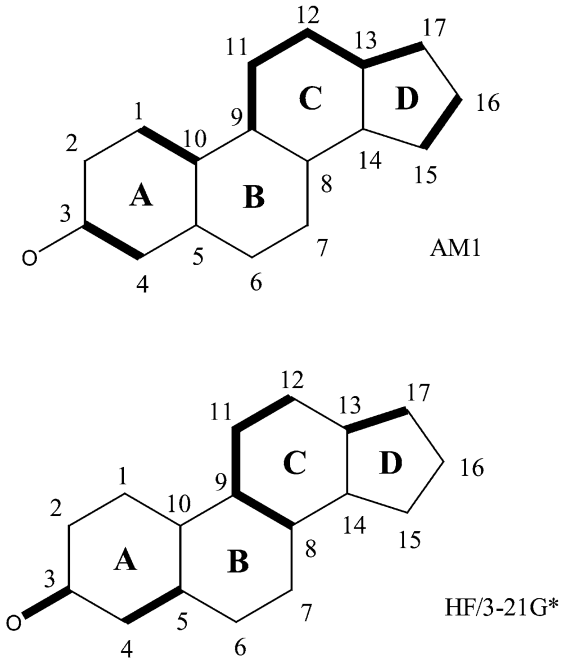

Figure 5. Bonds (in bold) selected by GA in the optimal PLS models for AM1 and HF/3-21G*.

π-population of heteroatoms at the position C_4 [31]. These results are all consistent with the regions of the molecule that the current method deems important in explaining the activity.

The results obtained using the GA for variable selection are given in Table 3 and the corresponding plot of the observed versus predicted steroid binding affinities for both models is given in Figure 4. It can be seen that employing variable selection leads to a marked increase in the quality of the model, as reflected by the superior regression statistics. A comparison with Figure 2 clearly reveals the improvement offered by the GA, in particular of steroid 31, 2α-methyl-9α-fluorocortisol, which is no longer an outlier. Most alternative methods [8, 14, 29, 47–51] predict steroid 31 to be an outlier however.

Figure 5 shows in bold the bonds selected by the GA. It is interesting to contrast the regions in the molecule picked out by the GA with those highlighted by the original model for both sets of bond lengths. In the AM1 model there is no particularly strong overlap between the bonds selected by the GA and the bonds highlighted whilst using all the descriptors. The GA fails to select either the carbon-oxygen bond or the carbon-carbon double bond, both of which are considered essential regions for steroid binding. However, the bonds selected by the GA at the HF/3-21G* level are in much broader agreement with the regions highlighted in the original model. Now both the carbon-oxygen bond and carbon-carbon double bond (C_4=C_5) contribute to the optimal model. We conclude that *GAs are best used to optimise the regression coefficients but should be employed in conjunction with the VIPs to provide greater interpretability of the models.*

Finally, we mention that, prompted by a referee comment, we expect the variations in bond length to correlate well with variations in vibrational motion seen at room temperature. This conjecture is based on an earlier study [23] proving the quite different QSAR of alkaline hydrolysis rate constants of polar esters can be predicted from relevant optimised bond lengths in the gas phase. In that paper we extensively referred to the work [52] of Collette, who showed the excellent correlation between such rate constants and his measured infrared gas-phase frequencies of the carbonyl bonds in 41 carboxylic esters.

Antibacterial activity of nitrofuran derivatives

Nitroheterocyclic drugs have enjoyed widespread application in medicinal chemistry [53], principally being used as antibacterial and anticancer agents. The compounds also possess mutagenic and carcinogenic activities [54]. It has been demonstrated that 5-

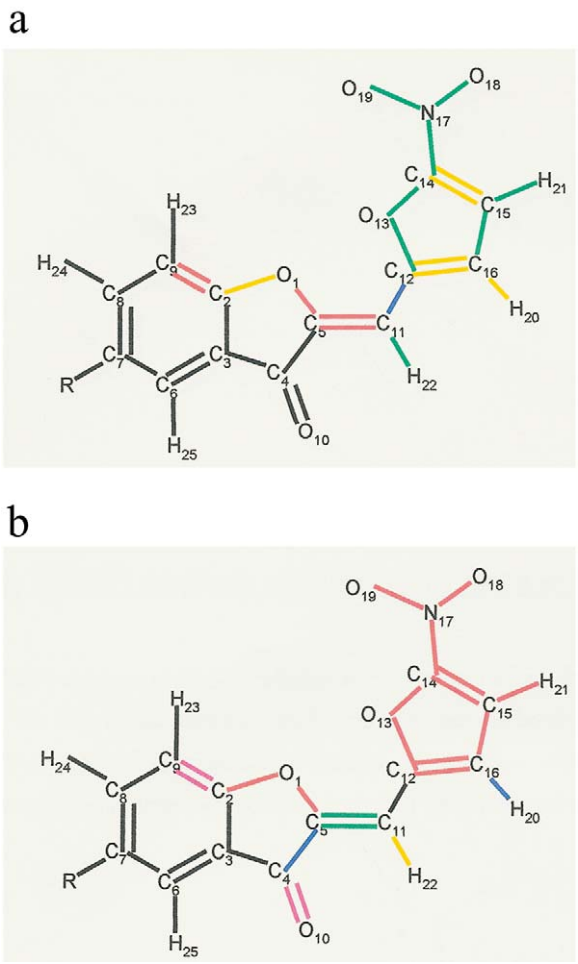

Figure 6. Colour-coded plot expressing the influence of the AM1 bond lengths in explaining the observed antibacterial activity of nitrofurans, for (a) *S. aureus* and (b) *C. crescentus*.

nitrofuran derivatives can damage proteins in addition to DNA [55], and despite the vast amount of experimental data on these compounds, the mechanism of action and the biological target are still unknown. In an effort to elucidate the structural features of this type of compounds that are important to antibacterial activity, Pires et al. [56] have performed QSAR analyses on three sets of nitrofuran derivatives [57], of which we study the first one. The structures of the common skeleton of the nine 5-R-substituted (Z)-2-(5-nitrofuran-2-ylmethylene)-3(*2H*)-benzofuranones or nitrofuran derivatives in short are shown in Figure 6, which also shows the atomic labeling.

The structural and activity data are given in Table 4. Here the activity is the negative logarithm of the 50% growth inhibitory activity, IC_{50}, or $\log(1/IC_{50})$, of the bacterial species *S. aureus* and of *C. crescentus*, each generating a different pattern of activity.

The reason we look at this (admittedly small) set is because Pires et al. [56] found that the single most important parameter in explaining the activity was expressed by any one of the electronic parameters. By using either the first reduction potential or any one of the substituent constants alone, the authors found that about 80% of the variance of growth inhibitory activity could be explained. In addition, lipophilicity was found to be unimportant in explaining growth inhibitory activity. Addition of the polarisability-related molar refractivity did lead to improved correlations, but upon examination of the regression coefficient for the parameter, the authors conclude that this term is not significant at the 95% confidence interval. In short, this is a dataset where electronic effects play a major role in determining the activity, and electronic effects are exactly the ones that QTMS is expected to be successful at (see Introduction).

For this set, however, we do *not* invoke a GA to select the variables (i.e. bond length) that provide the best cross-validated error or q^2 value. This provides a more stringent test of the predictive capacity of bond lengths. For the *S. aureus* activity a model (with two latent variables) was generated from the AM1 bond lengths with $r^2 = 0.92$ and $q^2 = 0.78$, and another model for *C. crescentus*, with $r^2 = 0.96$ and $q^2 = 0.90$, both passing the randomisation test.

Figure 6 shows the dramatic difference between the active site for the two types of bacteria. Pires et al. [56] state that the factor most likely to be important in regulating the activity of the nitrofurans is the stability/reactivity of the reduced intermediates. Since this is a global property concerning an intermediate step, it is difficult to associate this with a specific region in the neutral molecule. We hope that our finding may help researchers attempting to elucidate the as yet unknown mode of action of nitrofurans.

General discussion

We now comment on the advantages and potential limitations associated with the procedure as it stands. Currently the analysis has only been applied to predominately rigid systems exhibiting limited conformational flexibility. The extent to which conformational variation affects the QSARs is undetermined and will be investigated in due course. Secondly, the usual

caveats associated with QSAR, such as the need for a sufficiently large set of molecules with a common mode of action, are in operation. However, a rigorous statistical analysis should weed out any untrustworthy analyses. A third caution is the requirement of a common molecular skeleton between the structures, that is, the mapping of the atoms between molecules must be one-to-one and unambiguous.

An advantage, on the other hand, is that the method can be applied to any set of molecules that can be calculated via current computational schemes of quantum chemistry, thus circumventing the difficulties associated with 'classical' QSAR descriptors such as log P and the Hammett σ constants. A second advantage is the ability to independently predict the active site. As the majority of QSARs, particularly in the field of drug design, deal with novel compounds having unknown modes of action, the relevant parts ('active site' or 'reaction center') will often be unknown. A third advantage is that the application of the method does not require lengthy 3D molecular superpositions, nor decisions on grid size, spacing, orientation ('shift'), types of point charge, dielectric constants, truncation of electrostatic potential, etc. Lastly, given very feasible computing times, the descriptors are drawn from *ab initio* calculations that by their unparametrised nature provide solutions of the Schrödinger equation that are more direct and *in the limit* more realistic than those given by force fields. Of course, a HF/3-21G* calculation is by no means state of the art for molecules of the size of steroids, but given ever decreasing computing cost our method provides a straightforward avenue to contrast biological data with descriptors better connected with the underlying quantum reality. Naturally, our descriptors can feature in conjunction with other descriptors, such as log P, or be incorporated with non-linear models.

Finally, there is the ongoing debate about how elaborate or CPU time consuming a technique should be to produce a satisfactory QSAR, sometimes colourfully referred to as 'there are many expensive ways of counting carbons'. Before [58] we have shown that a superposition of electron densities is unnecessarily elaborate to predict Hammett σ-values. However, in the steroid case we query how much the precise knowledge of bond length is really required. For example, the $C_3{=}O_{18}$ bond length, which has a high VIP value, barely varies between steroids 4, 7 and 8. The large variation between double CO bonds on one hand and single CO bonds on the other is most likely the actual reason for the correct prediction of activity. Of course this example ignores the complexity of the full statistical analysis we have performed but it indicates that a hierarchical use of QSAR descriptors, starting with the ones of lowest computational cost, is recommendable.

Conclusions

In this study we report a new procedure for the localisation of molecular regions responsible for the observed biological activity in a congeneric series of molecules. The approach, based on the use of quantum-chemically obtained bond lengths as QSAR descriptors, was used to develop statistically satisfactory models describing the CBG binding affinity in the steroid data set. The notorious steroid 31, 2α-methyl-9α-fluorocortisol, is not found to be an outlier. The bond lengths obtained using semi-empirical AM1 calculations represent the best compromise between accuracy and computational cost. Genetic algorithm-based variable selection produces significant improvements in the quality of the models. Using convenient colour-coded plots the regions of the molecule considered important for binding activity were delineated. For the steroid set, the regions obtained correspond well with previous binding site specificity studies, but our suggestion for the nitrofurans remains open to confirmation. Despite obvious limitations we believe that the simplicity and general applicability of the method can provide an extra dimension to QSAR studies, particularly in view of the ease with which semi-empirical bond lengths can be obtained. Also, the ability to locate molecular regions responsible for a given biological activity could have great significance in the design of new biologically active compounds.

References

1. Hansch, C. and Leo, A. Exploring QSAR: Fundamentals and Applications in Chemistry and Biology. ACS, Washington, DC, 1995.
2. Karelson, M., Lobanov, V.S. and Katritzky, A.R., Chem. Rev., 96 (1996) 1027.
3. Karelson, M. Molecular Descriptors in QSAR/QSPR. Wiley-Interscience, New York, USA, 2000.
4. van de Waterbeemd, H. Methods and Principles in Medicinal Chemistry. VCH, Weinheim, Germany, 1995.
5. Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616.
6. Kubinyi, H., In Encyclopedia of Computational Chemistry (Schleyer, P.v.R., Ed.), Wiley, Chichester, UK, 1998, p. 2309.
7. Kim, K.H., In Molecular Similarity in Drug Design (Dean, P.M., Ed.), Chapman & Hall, London, 1995, p. 291.
8. Silverman, B.D. and Platt, D.E., J. Med. Chem., 39 (1996) 2129.

9. Klebe, G., Abraham, U. and Mietzner, T., J. Med. Chem., 37 (1994) 4130.
10. Cramer, R.D., III, Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
11. Carbo, R., Leyda, L. and Arnau, M., Int. J. Quant. Chem., 17 (1980) 1185.
12. Carbo-Dorca, R., Robert, D., Amat, L., Girones, X. and Besalu, E. Molecular Similarity in QSAR and Drug Design, Springer, Berlin, Germany, 2000.
13. Amat, L. and Carbo-Dorca, R., J. Med. Chem., 42 (1999) 5169.
14. Good, A.C., So, S.-S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.
15. Bader, R.F.W. Atoms in Molecules. A Quantum Theory, Oxford University Press, Oxford, UK, 1990.
16. Popelier, P.L.A. Atoms in Molecules. An Introduction, Pearson Education, London, UK, 2000.
17. Popelier, P.L.A., Aicken, F.M. and O'Brien, S.E. In Chemical Modelling: Applications and Theory (Hinchliffe, A., Ed.), Vol. 1 Royal Society of Chemistry Specialist Periodical Report, Ch. 3, 2000, pp. 143–198.
18. Popelier, P.L.A. and Smith, P.J. In Chemical Modelling: Applications and Theory (Hinchliffe, A., Ed.), Vol. 2, Royal Society of Chemistry Specialist Periodical Report, Ch. 8, 2002, pp. 391–448.
19. Popelier, P.L.A., J. Phys. Chem. A, 103 (1999) 2883.
20. Popelier, P.L.A., Chaudry, U. and Smith, P.J., J. Chem. Soc., Perkin Trans. 2 (2002) 1231.
21. O'Brien, S.E. and Popelier, P.L.A. Quantum Molecular Similarity: Use of Atoms in Molecules derived quantities as QSAR variables. ECCOMAS, Barcelona, Spain, 2000.
22. O'Brien, S.E. and Popelier, P.L.A., J. Chem. Soc., Perkin Trans. 2 (2002) 478.
23. Chaudry, U.A. and Popelier, P.L.A., J. Phys. Chem. A, 107 (2003) 4578.
24. Chaudry, U.A. and Popelier, P.L.A., J. Org. Chem., 69 (2004) 233.
25. O'Brien, S.E. and Popelier, P.L.A., Can. J. Chem., 77 (1999) 28.
26. Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
27. So, S.-S. and Karplus, M., J. Med. Chem., 40 (1997) 4360.
28. Robinson, D.D., Winn, P.J., Lyne, P.D. and Richards, W.G., J. Med. Chem., 42 (1999) 573.
29. Robert, D., Amat, L. and Carbo-Dorca, R., J. Chem. Inf. Comput. Sci., 39 (1999) 333.
30. Maw, H.H. and Hall, L.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1248.
31. Coats, E.A. Perspect. Drug Discov. Design, 12 (1998) 199.
32. Schaftenaar, G. and Noordik, J.H., J. Comput.-Aided Mol. Design, 14 (2000) 123.
33. GAUSSIAN98. Gaussian 98, Revision A.7, M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, Jr., R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, A.G. Baboul, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B. Johnson, W. Chen, M.W. Wong, J.L. Andres, C. Gonzalez, M. Head-Gordon, E.S. Replogle, and J.A. Pople, Gaussian, Inc., Pittsburgh, PA, USA, 1998.
34. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
35. Foresman, J.B. and Frisch, A. Exploring Chemistry with Electronic Structure Methods, Gaussian Inc., Pittsburgh, PA, USA, 1996.
36. MORPHY98. A program written by P.L.A. Popelier with a contribution from R.G.A. Bone, UMIST, Manchester, UK (1998). http://morphy.ch.umist.ac.uk/).
37. Wold, S., Sjostrom, M. and Eriksson, L. In: Encyclopedia of Computational Chemistry (Schleyer, P., Ed.), Wiley, Chichester, UK, 1998, p. 2006.
38. Wold, S., Kettaneh, N. and Tjessem, K., J. Chemometrics, 10 (1996) 463.
39. Holland, J.H. Adaption in Natural and Artificial Systems, MIT Press, Cambridge, MA, USA, 1992.
40. UMETRICS. info@umetrics.com: www.umetrics.com, 2002.
41. Livingstone, D.J. Data Analysis for Chemists, Oxford University Press, Oxford, UK, 1995.
42. Wold, S., In van de Waterbeemd, H. (Ed.) Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 195.
43. O'Brien, S.E. Ph.D. Thesis, Department of Chemistry, UMIST, Manchester, UK, 2000.
44. Wise, B.M. and Gallagher, N.B. Eigenvector Research, Manson, WA, USA, 2003.
45. Mickelson, K.E., Forsthoefel, J. and Westphal, U., Biochemistry, 20 (1981) 6211.
46. Amat, L., Besalu, E. and Carbo-Dorca, R., J. Chem. Inf. Comput. Sci., 41 (2001) 978.
47. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H., J. Comput.-Aided Mol. Design, 10 (1996) 513.
48. Lobato, M., Amat, L., Besalu, E. and Carbo-Dorca, R., Quant. Struct.-Act. Relat., 16 (1997) 465.
49. Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2091.
50. Tominaga, Y. and Fujiwara, I., J. Chem. Inf. Comput. Sci., 37 (1997) 1152.
51. Schnitker, J., Gopalaswamy, R. and Crippen, G.M., J. Comput.-Aided Mol. Design, 11 (1997) 93.
52. Collette, T.W., Environ. Sci. Techn., 24 (1990) 1671.
53. Debnath, A.K., Hansch, C., Kim, K.S. and Martin, Y.C., J. Med. Chem., 36 (1993) 1007.
54. McCalla, D.R., Environ. Mutagen., 5 (1983) 745.
55. Peterson, F.J., Mason, R.P., Hovsepian, J. and Holtzman, J.L., J. Biol. Chem., 254 (1979) 4009.
56. Pires, J.R., Saito, C., Gomes, S.L., Giesbrecht, A.M. and doAmaral, A.T., J. Med. Chem., 44 (2001) 3673.
57. Cheng, E., Haiduke, R.L.A., Pires, R., Ishiki, H., Bruns, R.E. and do Amaral, A.T. In: EuroQSAR2002: Designing Drugs and Crop Protectants: processes, problems and solutions (Ford, M., Livingstone, D.J., Dearden, J., van de Waterbeemd, H., Eds.), Blackwell, Oxford, UK, 2003, p. 166.
58. O'Brien, S.E. and Popelier, P.L.A., J. Chem. Inf. Comput. Sci., 41 (2001) 764.