Journal of Computer-Aided Molecular Design **18:** 709–718, 2004.
DOI 10.1007/s10822-004-6815-7

Quantitative structure-activity relationships of mutagenic activity from quantum topological descriptors: triazenes and halogenated hydroxyfuranones (mutagen-X) derivatives

P.L.A. Popelier*, P.J. Smith & U.A. Chaudry
School of Chemistry, University of Manchester, Manchester M60 1QD, UK

Received 9 August 2004; accepted in revised form 25 November 2004

Key words: ab initio, active center, mutagens, triazenes, hydroxyfuranones (MX), QSAR

Summary

The mutagenic activity of 23 triazenes and, in a different set, of 24 halogenated hydroxyfuranones (MX derivatives) is quantitatively related to new features of contemporary molecular wave functions. Nowadays affordable computers are powerful enough to rapidly generate geometry-optimised *ab initio* wave functions at HF/3-21G*, HF/6-31G* and B3LYP/6-311 + G(2d,p) level for all molecules. The bonds of a common molecular skeleton are described by their *ab initio* bond lengths and local properties provided by the theory of quantum chemical topology (QCT). The chemometric analysis involves two types: one to generate a statistically validated quantitative model, and one to isolate the active center. In the former a genetic algorithm (GA) selects bond descriptors in order to optimise the cross-validation error, q^2, followed by a full partial least squares (PLS) analysis, which also yields randomisation statistics. In the latter type principal components (PCs) are constructed from the original bond descriptors and their variables important to the projection (VIPs) are plotted in a histogram. This analysis suggests a preferred mechanistic pathway for the initial hydroxylation of the triazenes, an issue that has remained ambiguous so far. In the case of the hydroxyfuranones the proposed method aids the elucidation of a mechanistic ambivalence.

Introduction

Mutagens are agents that cause mutation or genetic change. For example, purines or pyrimidines that are not found in natural DNA can be mutagenic if they are more likely than natural DNA to mispair [1]. Many known carcinogens are either mutagens or are converted to mutagens by enzymes. The latter process is the basis for the well-known Ames test [2], in which the microsomes of liver cells convert carcinogens to mutagens. This inexpensive and reproducible test selects for newly mutated bacteria and counts their colonies. The presence of mutant bacteria indicates that the compound added to the suspension of liver cells, suspected carcinogen and original bacteria is mutagenic. Since there is a correlation between activity, in the Ames test and tumor-forming activity, the data resulting from the Ames test are relevant in elucidating the chemical mechanisms by which mutagenic chemicals interact with DNA.

In this paper we apply a newly developed method, called quantum topological molecular similarity (QTMS) [3], to two sets of mutagenic compounds: dimethyl heteroaromatic triazenes [4] and halogenated hydroxyfuranones [5]. The latter compounds are analogues of Mutagen-X (or MX), one of the strongest bacterial mutagens ever tested. Because the mechanisms governing the two sets of mutagenic compounds are different, we treat each set separately.

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

QTMS is essentially a method that uses descriptors from quantum chemical topology (QCT) [6–9] to compare and differentiate molecules, such that insight into their activity can be obtained and predictions of their activity can be made. Since its inception [10], QTMS has been developed in the framework of quantitative structure activity/property relationships (QSAR/QSPR) but it is not restricted to it. QTMS has delivered novel QSARs both in a medicinal and ecological context. As examples of the medicinal QSARs we mention the anti-tumor activity of (E)-1-phenyl-but-3-en-ones [11], the antibacterial activity of nitrofuran derivatives and the steroid binding affinity of CBG [12]. Successful ecological QSARs include a study on the toxicity and biodegradability of *p*-substituted phenols and ^{13}C NMR chemical shifts in *p*- and *m*-substituted benzonitriles [13], the toxicity of polychlorinated dibenzo-p-dioxins (PCDDs) [14] and the prediction of hydrolysis rate constants of polar esters [15]. The estimation of pK_a of carboxylic acids, anilines and phenols [16] and prediction of σ_p, σ_m, σ_I and σ_p^0 parameters of mono-[10] and polysubstituted benzoic acids, phenylacetic acids and bicyclo carboxylic acids [14] feature as examples of physical organic properties.

In this contribution we first show how QTMS provides reliable and cross-validated predictions for mutagenicity in two well-documented cases, that is, a set of substituted triazenes and a set of substituted hydroxyfuranones. Then we demonstrate how QTMS sheds light on the mechanism of action. In the case of the triazenes we can invoke QTMS to resolve a mechanistic issue in favour of one of two proposed alternative pathways. In the case of the hydroxyfuranones QTMS may help resolving a mechanistic ambivalence between nucleophilic addition to the unsaturated ring and a one-electron transfer process.

Features of quantum topological molecular similarity (QTMS)

Computational details are provided in the next section. Here we briefly review the background of QTMS, which essentially proposes a chemometric analysis of QCT descriptors, as described in ref. [3]. In this paper the use of a Genetic Algorithm (GA) for variable selection appears as a new feature, but we also discuss some results obtained without it.

Since a didactic review of the topology of the electron density, denoted by ρ, can be found elsewhere [9] we only mention salient points here. Critical points (CP) are points in real 3D space where the gradient of ρ vanishes. Of the four possible types of CPs only one type is of interest here. It is called the bond critical point (BCP) because it appears between bonded nuclei, loosely speaking. The Hessian matrix evaluated at the BCP is crucial to confirm that a CP is indeed a BCP. This symmetric three-by-three matrix contains second derivatives as elements, that is $\partial^2\rho/\partial q_i \partial q_j$, where q can be x, y or z. Two out of three eigenvalues of the Hessian, denoted by λ_1, λ_2 and λ_3 ($\lambda_1 < \lambda_2 < \lambda_3$) must be negative for the CP to have the correct signature associated with a BCP. The negative eigenvalues λ_1 and λ_2 are curvatures perpendicular to the bond, while the positive eigenvalue λ_3 measures the curvature along the bond.

It is important to recognise that the BCP is a point that, once localised [17, 18], serves as a quantum chemical signature for a bond via the properties evaluated at its position. For example, it has been proposed [19] to relate the electron density evaluated at the BCP, ρ_b, to bond order via an exponential relationship. This interpretation seems to be successful for carbon–carbon bonds only [20], which does not invalidate the fact that ρ_b still functions as a quantum chemical descriptor retrieved from a modern *ab initio* wave function; only its meaning for a general (non-CC) bond may not be that straightforward. However, a multiple linear relationship of the type $a + b\rho_b + c\lambda_3 + d(\lambda_1 + \lambda_2)$ has been recommended [20] to predict a topological bond order [21] more rapidly than via direct evaluation using atomic overlap matrices. In that work [20], ρ_b and λ_3 were interpreted as measures of σ character, whilst $(\lambda_1 + \lambda_2)$ measures the degree of π character. In this work we employ an alternative measure, called the ellipticity, defined as $\varepsilon = (\lambda_1/\lambda_2) - 1$. The ellipticity has traditionally been used as a measure for π character for homopolar bonds but is a measure of structural stability [22] in a more rigorous context. In spite of its variation along polar bonds [23] it can safely be used when evaluated at the BCP because of the small changes in its position for a given bond going from one substituted compound to another. Another useful BCP property is the Laplacian, $\nabla^2\rho$, which can be defined as $\lambda_1 + \lambda_2 + \lambda_3$. If negative it indicates that electronic

charge is locally concentrated. It is often used as a simple measure of covalent character. We also considered a kinetic energy density evaluated at a BCP as a further BCP property to be included in the list of QCT descriptors. This density, denoted by $K(\mathbf{r})$, is defined as

$$K(\mathbf{r}) = -\frac{1}{4}N \int d\tau' [\psi^{\star}\nabla^2\psi + \psi\nabla^2\psi^{\star}] \qquad (1)$$

where ψ is the many-electron wave function and $\int d\tau'$ denotes an integration over the spin coordinates of all N electrons except one. Interpretting $K(\mathbf{r})$ in chemical terms is not straightforward although useful formulae describing its link to the Laplacian and the more 'classical' kinetic energy $G(\mathbf{r})$ can be found elsewhere [24]. Finally, we decided to incorporate the equilibrium bond length, R_e, as a fifth descriptor for a given bond. As detailed below all *ab initio* wave functions correspond to optimised geometries. This descriptor is not strictly a BCP property but, for sake of consistency, it can somewhat artificially be turned into one. For that purpose, one regards R_e as a sum of two distances, i.e. the distance between the BCP and one nucleus and the distance between the same BCP and the other nucleus, neglecting any deviation from a straight line the bond path [9] may exhibit. In summary, a vector with five linearly independent components, namely (ρ_b, $\nabla^2\rho$, ε, K, R_e), describes each bond.

Computational method

After the program MOLDEN [25] has estimated an initial molecular geometry for each compound, the *ab initio* programme GAUSSIAN98 [26] optimises the geometries at four different levels of theory. Optimisation refers to finding a local minimum in the potential energy surface (at 0 K), which describes the molecular electronic energy as a function of the nuclear coordinates. We used, in succession, AM1 [27], HF/3-21G*, HF/6-31G* and B3LYP/6-311 + G(2d,p) [28, 29], passing on the optimised geometry of each level as a starting geometry for the next. The issue of conformation was not explicitly addressed but is expected to be minor for the current set of molecules. Previous QTMS studies have always successfully worked for local minima, without considering higher energy conformations. For convenience and consistency with our previous QTMS studies we refer to these levels of theory as *A, B, C* and *E*, respectively. Once the wave function files have been generated they are read by (a local version of) the program MORPHY98 [30]. This program computes the BCP properties, extracts the required bond lengths and exports them into a format convenient for subsequent statistical analysis. The CPU time required to compute the BCP properties is marginal compared to that to generate the wave functions. Note that we are not using QCT integrated properties, which are known to consume a lot of CPU time.

We point out that the current version of QTMS still requires a 1–1 mapping between bonds of different substituted compounds. This prerequisite prompts the identification of a common molecular skeleton, which is a subset of bonds that can unambiguously be identified in each compound participating in the QSAR. For example, in the triazene QSAR the common molecular skeleton consists of 7 bonds as shown in Figure 3. For a more detailed discussion and a relaxation of the common skeleton condition the reader is referred to an earlier paper [31] in the QTMS series.

In the chemometric stage we construct a model via Partial Least Squares (PLS) [32, 33] in conjunction with a GA-based [34] variable selector. The program SIMCA-P [35] yields the latent variables (LV) and yields four statistics. The first statistic is the correlation coefficient, r^2, used in conjunction with the second statistic, the cross-validated r^2, denoted by q^2. The latter is dependent on the PRESS score [36], calculated here by leaving out one seventh of the data. Some groups prefer to split the data into a training set and a test set. Although this is a stringent test, the decision as to which compound belongs to the training set or the test set is not unique. We believe that our current leave-*group*-out [37] style scheme avoids the well-known pitfall of the leave-*one*-out procedure. The final two statistics, obtained via a 'Y-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'. This is mathematically expressed by stating that a model is considered valid [32] if $r^2_{int} < 0.4$ and if $q^2_{int} < 0.05$.

In order to obtain the best possible r^2 and q^2 statistics we invoked a GA to optimise the number

of descriptors for use in a subsequent PLS analysis. A new PLS model is then obtained using the descriptors selected by the GA. We employed a GA as implemented in the MATLAB routine *genalg.m* from the PLS Toolbox [38]. The GA introduced a population of 256 randomly selected models and the fitness function was the cross-validation error q^2. The mutation rate was set at 0.003 and the maximum number of generations at 200.

In a second analysis we use *all* original descriptors once more, now without GA selection. It is appropriate to examine the so-called variables important to the Projection (VIPs) [39], also calculated by SIMCA-P, in order to interpret the model obtained. The VIPs give the relative importance of each descriptor contributing to the model [40]. Descriptors with higher VIP scores are considered more relevant in explaining the activity. In order to bundle all descriptors associated with a given bond into one descriptor (or 'X' variable) it is convenient to construct Principal Components (PCs) from them. The program SPSS [41] performs this data (unsupervised) compression and guarantees that maximally localised BCP information. In this work the PCs with the highest VIP value will be plotted as histograms showing the VIP value. 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. Variables with a VIP value of less than one are considered unimportant to the model [32] and can be discarded.

Results and discussion

Triazenes

The 23 heteroaromatic triazenes [4] (**1–23**) investigated in this work are listed in Table 1, with the corresponding structural diagrams shown in Figure 1. Note that all compounds have at least one methyl group attached to the unique three-coordinate nitrogen in the N=NN group. The mutagenicity of compounds **1–18**, which are all substituted 1-methyl-1-alkyl-3-phenyltriazenes, was measured by Venger et al. [42] in the Ames bacterium *Salmonella typhimurium* strain TA92 to which the S9 microsome preparation had been added. The mutagenicity of the remaining five compounds (**19–23**), which are heterocyclic 3,3-dimethyltriazenes, was measured by Shusterman et al. [4] according to the same method. Mutagenicity is quantitatively expressed as log(1/*C*) where *C* is the molar concentration of triazene that causes 30 mutations above background per 10^8 TA92 bacteria. The higher the value of log(1/*C*) the more mutagenic the compound. The data set under study spans an activity range of 5.55 log unit corresponding to a 350,000 fold variation. It is convenient to already note here that it is actually R^+ (or CH_3^+ for **19–23**) that reacts with DNA. Note that QTMS does not predict logP values; hence these values were adopted from elsewhere [4, 42] and are repeated for convenience in Table 1. QTMS does not generate steric descriptors either. In fact, over the years it has become clear that QTMS only captures the electronic effects to the extent that when QTMS delivers a poor QSAR one can safely conclude that steric and lipophilicity

Table 1. Measured mutagenic activity (log(1/*C*)) and log *P* values for triazene derivatives.

No.	X[a]	R[a]	log 1/*C*	log *P*
1	4-$CONH_2$	*t*-Butyl	3.83	2.61
2	3,5-CN	CH_3	3.46	2.18
3	4-SO_2NH_2	CH_3	3.49	0.98
4	3-$CONH_2$	CH_3	3.51	1.21
5	4-$CONH_2$	CH_3	4.04	1.20
6	4-$CONH_2$	Allyl	4.16	2.09
7	3-$NHCONH_2$	CH_3	4.19	1.29
8	4-CN	CH_3	4.43	2.39
9	4-$COCH_3$	CH_3	4.47	2.27
10	H	CH_3	5.32	2.59
11	4-$CONH_2$	*n*-Butyl	5.41	2.46
12	4-$NHCONH_2$	CH_3	5.59	1.25
13	4-$NHCOCH_3$	CH_3	5.83	1.54
14	4-CF_3	CH_3	5.99	3.70
15	3-CH_3	CH_3	6.44	2.85
16	4-Cl	CH_3	6.48	3.33
17	4-CH_3	CH_3	7.00	2.93
18	4-C_6H_5	CH_3	7.67	4.40
19	DTIC[b]	CH_3	3.00	−0.24
20	2-1′, 3′-Thiazolidine		5.90	1.27
21	2-Dibenzofuran		8.55	4.55
22	5-Indazole		6.53	2.36
23	2-Benzimidazole		6.48	1.38

[a]The substituents X and R refer to compounds 1 to 18 in Figure 1.
[b]DTIC = 5-(3,3-*d*imethyl-1-*t*riazeno) *i*midazole-4-*c*arboxamide.

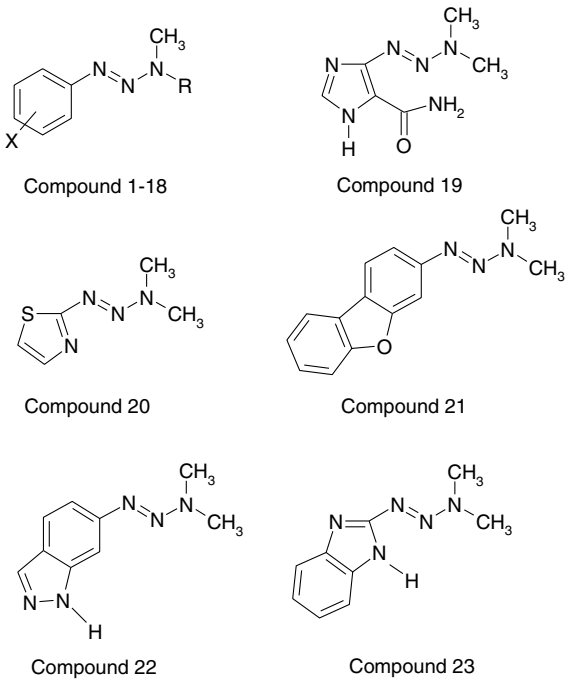

Figure 1. Schematic structures of triazene derivatives. The substituents *R* and *X* of compounds **1** to **18** are specified in Table 1.

descriptors are not relevant to explain the activity. Hence, at this stage of development QTMS relies on classical methods for log *P* and steric descriptors. The latter did not feature in the paper by Shusterman et al. [42], which is why only log *P* was included in our analysis. Before discussing the results of the current QTMS analysis we highlight relevant findings of other researchers who worked on triazenes.

In their study Venger and co-workers [42] proposed a QSAR featuring log *P* (where *P* is the octanol–water partition coefficient) and σ^+, the Hammett substituent constant for reactions involving through-resonance. Their equation, encompassing compounds **2–18** and excluding compound **1** because of poor prediction, yielded $r^2 = 0.949$. Although the antitumor drug DTIC (**19**) continued to be widely used in clinical antitumor studies Venger et al. could not include it in their QSAR equations because they lacked the appropriate σ constant. No cross-validation measure (i.e. q^2 predictivity) or randomisation test statistics were provided.

About 10 years later Shusterman et al. [43] calculated the electronic structure of triazenes via the semiempirical MNDO method. Their data put them in a position to replace the Hammett constant σ^+ by orbital related quantities. More specifically, they used the energy of a triazene's HOMO and the HOMO electron density associated with the nitrogen carrying the two alkyl groups (i.e. the unique three-coordinate nitrogen). The HOMO energy is global in that it involves all atoms, whereas the other property is local, because it only involves only one atom. We emphasise that the QTMS descriptors are all local since each of them is clearly associated with a particular bond. In a second study Shusterman et al. [4] extended their approach beyond the set of phenyltriazenes. The QTMS method shares with their approach the advantage of being able to include compounds other than phenyltriazenes. Indeed, σ^+ is only available for phenyltriazenes and hence QSARs based on it can only treat a limited range of mutagen structure and activity. Relying on contemporary *ab initio* data QTMS is independent of the availability of any type of Hammett constant.

Triazenes do not appear to be carcinogenic or mutagenic until they have been activated by microsomes. The microsomal cytochrome P-450 hydroxylates the *N*-methyl group, which is present in all compounds. It has been postulated that, after this initial hydroxylation, triazenes undergo a number of reactions that ultimately lead to a carbonium ion. This ion, R^+, is the active species attacking DNA in mutagenesis [42]. An important question is whether a structural feature of R^+ itself should be linked with mutagenic activity. Shusterman et al. [4] proposed an answer by introducing the plausible hypothesis that it is actually the ease of initial triazene activation (i.e. hydroxylation by P-450) that governs mutagenicity, rather than the rate of DNA alkylation by R^+. In other words, the relative activity of phenyltriazenes correlates with their initial activation rate. This assumption is based on their observation that the factors correlating with increased mutagenicity are the same factors favouring hydroxylation [44].

Figure 2 zooms in on the triazene hydroxylation. The details of the hydroxylation mechanism are still obscure but the two most likely alternatives are shown. Path *A* involves transfer of one electron from the triazene to the enzyme cytochrome P-450. The resulting radical cation then loses a proton to give a triazene-substituted methyl

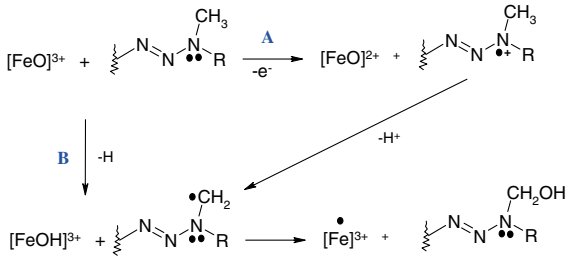

Figure 2. Proposed alternative mechanistic paths (*A* and *B*) of triazene hydroxylation by the enzyme cytochrome P-450, which contains the Fe group.

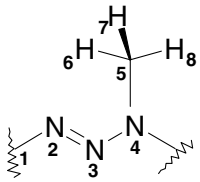

Figure 3. Common skeleton of all triazenes (**1–23**), consisting of seven bonds, referred to throughout the text via the atomic numerical labels.

radical. In the alternative path *B* this radical is formed in one step by direct abstraction of a hydrogen atom. Previously published QSAR equations [4] were unable to distinguish these two paths and suggest a preferred route. Pires et al. [45] extended the concept of frontier reactivity index, proposed by Fukui et al. [46], to groups of atoms. Their descriptors reflect the role of frontier electron density in promoting electrophilic attack on the triazene group and are hence expected to help in elucidating the reaction scheme of Figure 2. They generated a set of equations for the same QSAR set as in this work and obtained respectable r^2 values of around 0.87. However, in contrast to QTMS, their method was not capable of indicating a preferential path for the mechanism of triazene hydroxylation.

Figure 3 shows the molecular skeleton consisting of seven bonds, common to all triazene compounds. As expected this common molecular skeleton contains the NNN group and the methyl group, which is invariably present and attached to the unique tri-coordinated nitrogen (N_4). The skeleton is obviously that part of a triazene molecule that undergoes hydroxylation and is thus expected to model the limiting step in the mutagenic process. It is also large enough to cover the difference in mechanistic pathways *A* and *B*.

Table 2. Results of GA-PLS analysis of triazenes at four levels of theory.

Level	LV[a]	r^2	q^2
AM1 (A)	1	0.451	0.116
HF/3-21G* (B)	2	0.808	0.461
HF/6-31G* (C)	3	0.818	0.671
B3LYP/6-311 + (2d,p) (E)	3	0.859	0.741

[a]Number of latent variables.

Table 2 shows a summary of the models obtained for all 23 triazenes at four different levels of theory. Only seven descriptors are included in the semi-empirical method because AM1 wave functions do not yield BCPs. This is due to the absence of core orbitals [47]. On the other hand, the *ab initio* levels (*B, C* and *E*) furnish 35 descriptors (five for each of the seven bonds). Note that log *P* was added as an extra descriptor in each model. Since the q^2 drops dramatically without log *P* it is a vital part of the SAR. However no valid model can be obtained with log *P* alone, i.e. without the electronic descriptors of QTMS. Table 2 makes clear that the higher level of theory, the higher the predictive power of the model. In general more expensive levels of theory cannot always be justified, unlike in this case.

In Figure 4 we show a plot of the observed versus computed mutagenic activity (log(1/*C*)) for the best model (level E) including all triazenes

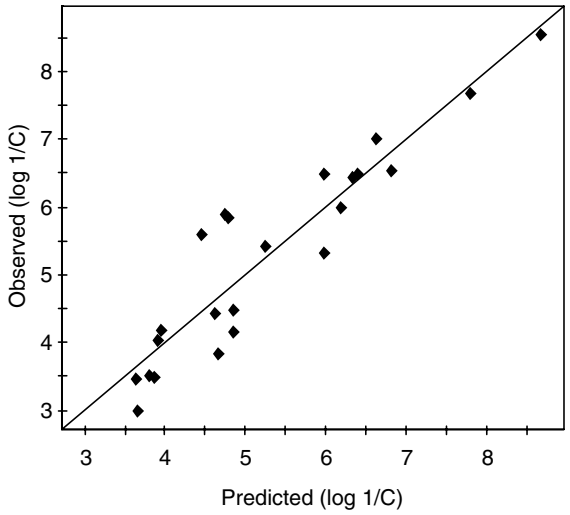

Figure 4. Observed versus computed mutagenic activity (log (1/*C*)) for the best model (level E) including all triazenes (**1–23**).

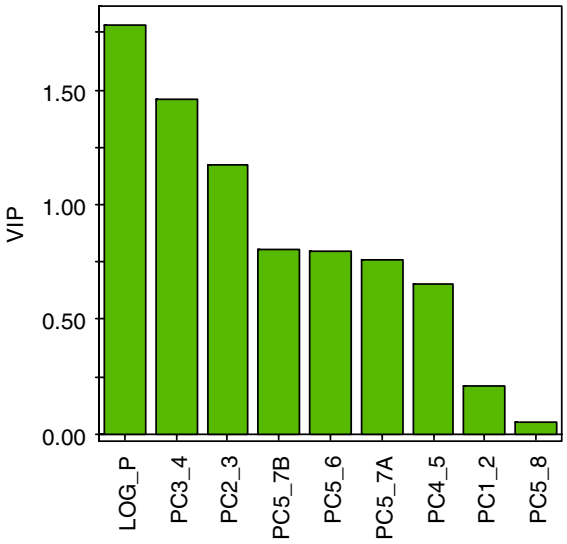

Figure 5. Variables Important to the Projection (VIPs) values for all PCs describing the bond properties of the common molecular skeleton (and log *P*), for the best model (level E) obtained including all triazenes (**1–23**).

(**1–23**). No significant outliers are detected, in contrast to previous work [4, 45] where compounds **1** and **20** had to be excluded. In that work the bulky nature of the *tert*-butyl group in **1** was called upon in order to explain the discrepancy between prediction and measurement. The models of those authors lack steric descriptors and the low measured activity has been attributed to steric hindrance [42]. Our model suggests there is no need for this explanation. Compound **20**, thiazolidine, contains a sulphur atom in a heterocyclic five-membered ring. Previous work, which relied on semi-empirical calculations, could not decide if this five-membered ring was poorly modelled or whether the high activity of thiazolidine was due to an additional mechanism for inducing mutations. Given that **20** is not an outlier for the *ab initio* model *E* we rule in favour of the former explanation.

Figure 5 shows a histogram of VIP values for all PCs describing the bond properties of the common molecular skeleton and log *P*. This plot corresponds to the best model, obtained at level *E*, and includes all triazenes (**1–23**). The most important PC, after the log P variable, refers to the N_3–N_4 bond. The next most important bond is the N_2=N_3 bond. After a substantial decrease in VIP value, PCs referring to the methyl group centered on C_5 follow. In summary, the N=N—N group is considerably more important in explaining the mutagenic activity than the omnipresent methyl group. In line with many other QTMS case studies [3, 10–12, 15, 16, 31, 48] we can conclude that the active site of the molecule is the NNN group rather than the methyl group. As a consequence pathway *A* is favoured over pathway *B*. From Figure 2 it is clear that the abstraction of a hydrogen atom in pathway *B* brings about an immediate change in the methyl group. If pathway *B* were correct then this change should express itself in high VIP values for methyl C—H bonds, just as the O—H bond has a high VIP value in the QTMS study [3] of the acidity of substituted benzoic acids.

Mutagen-X (MX) derivatives

MX or 3-chloro-4-(dichloromethyl)-5-hydroxy-2(5H)-furanone is one of the strongest bacterial mutagens ever tested. It is listed as the first compound in Table 3 where its measured [49] mutagenic activity is expressed as ln(TA100), referring to the Ames *Salmonella typhimurium* TA100 assay. MX and the 23 derivatives (Figure 6) incorporated in this QTMS study span 12.3 natural logarithm units or a range of TA100 activities of 1 to 210,000. The MX derivatives are ranked according to decreasing activity. The members of the MX family have in common with the well-known *cis*-platinum compounds that they do not need microsomal activation but alkylate the DNA directly to cause mutation [5]. Thus they operate in a different way to the triazenes, whose microsomal hydroxylation was identified as a key step in their mutagenic activity. The MX derivative set is unique [50] amongst mutagens in that log *P* plays no role [5], thereby rendering the set potentially more suitable for a QTMS analysis. There is evidence [50] that the lack of importance of hydrophobicity is a feature of direct-acting mutagens, i.e. those agents that do not need microsomal activation. This puts a MX derivative on a par with the R^+ moiety that appeared in the triazene study above as the actual agent attacking DNA. Thus it seems evident that the mutagenicity of MX is mainly a manifestation of its electron-accepting ability [5]. This sole dependence of mutagenicity on reductive properties of the molecules is very rare.

In the first of a series of studies Tupparainen et al. [51] found significant correlations between

Table 3. Measured mutagenic activity data (ln(TA100)) for the MX analogues.

Molecule	R_1	R_2	R_3	ln(TA100)	Molecule	R_1	R_2	R_3	ln(TA100)
1	$CHCl_2$	Cl	OH	8.748	13	CH_2Br	Cl	H	1.374
2	$CHCl_2$	Cl	OCH_3	8.648	14	CH_2Cl	Br	H	1.371
3	$CHBr_2$	Cl	OH	8.607	15	CH_2Cl	H	H	1.351
4	$CHBr_2$	Br	OH	7.966	16	Cl	Cl	OCH_3	0.993
5	CH_2Cl	Cl	OH	6.361	17	Cl	Cl	OCH_2CH_3	0.742
6	CH_2Br	Br	OH	6.04	18	Cl	Cl	OH	0.405
7	$CHBr_2$	Cl	H	5.198	19	CH_3	Br	OH	0.405
8	$CHCl_2$	Cl	H	5.176	20	Cl	Cl	H	0.113
9	$CHBr_2$	Br	H	4.86	21	H	Cl	OCH_2CH_3	−0.223
10	Cl	Cl	OH	4.094	22	CH_2Cl	H	H	−1.187
11	CH_2Br	Br	H	2.109	23	H	Cl	OH	−1.603
12	CH_2Cl	Cl	H	1.593	24	CH_3	H	OH	−3.507

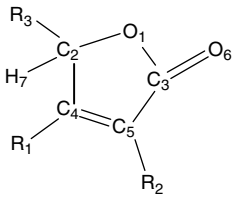

Figure 6. Schematic structures of MX derivatives. The substituents R_1, R_2 and R_3 are specified in Table 3.

the TA100 mutagenicity and the LUMO energy (E_{LUMO}), electron affinity, LUMO electron density and partial charge at the α carbon (C_5), all computed at AM1 level. Many of these parameters are collinear however. Steric and hydrophobic factors were found to be of minor importance. The subsequent formulation of a high-correlation QSAR equation [5] ($r^2 = 0.92$) containing only E_{LUMO} made the authors support the hypothesis that mutagenesis involves an electron transfer from a nucleophilic DNA base to the lowest empty molecular orbital of the MX compound. The lower the E_{LUMO} (more negative) value, the higher the mutagenic activity. We independently confirmed a good correlation with HF/3-21G* LUMO energies ($r^2 = 0.81$, $q^2 = 0.80$), a very poor one with HOMO–LUMO gaps and no model with HOMO energies. In spite of this encouraging correlation with E_{LUMO} the site of attack remains undecided. Attention focuses naturally on the electrophilic π system consisting of the C=C—C=O group. However, the physical meaning of an E_{LUMO} QSAR equation continues to be mechanistically ambivalent. Indeed, a correlation with E_{LUMO} cannot reveal the site of attack, for example. Two alternatives proposed before [50] are of interest to QTMS's capability because they implicate a localisation of the reactivity. In one mechanism, the nucleophilic attack and consequently the reaction with DNA would mainly occur at the β carbon (C_4). In another mechanism, the observed correlations between the mutagenicity and frontier-orbital electron densities suggests that the reaction center is near the α carbon (C_5) [5, 52], which is in good agreement with a CoMFA study [53].

Table 4 lists the results of the QTMS GA-PLS analysis of all MX derivatives at four levels of theory. The HF/6-31G* (level C) demonstrates the best predictive power. The concurrence of predicted and measured ln(TA100) values can be seen in Figure 7. In Figure 8 we show the VIPs values for GA-selected BCP properties of the bonds in the common molecular skeleton for the best model (level C), obtained including all MX derivatives. The four most important BCP properties all involve the keto carbon C_3. The fact that the highest VIP value is associated with the O_1—C_3 bond hints at a mechanism proposed earlier [50], suggesting

Table 4. Results of GA-PLS analysis of MX derivatives at four levels of theory.

Level	LV[a]	r^2	q^2
AM1 (A)	1	0.533	0.396
HF/3-21G* (B)	2	0.763	0.528
HF/6-31G* (C)	3	0.832	0.720
B3LYP/6-311 + (2d,p) (E)	1	0.752	0.672

[a]Number of latent variables.

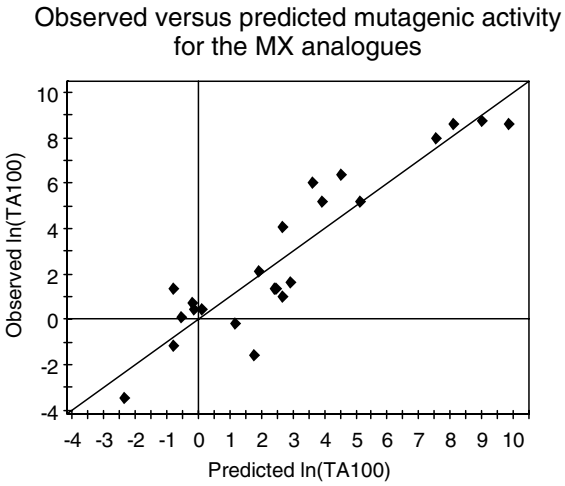

Figure 7. Observed versus computed mutagenic activity (ln(TA100)) for the best model (level C) including all MX derivatives.

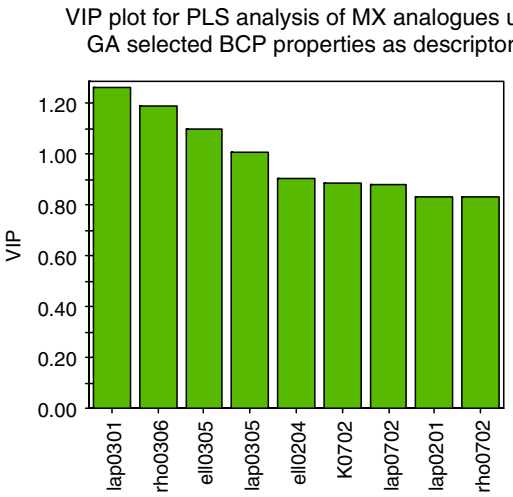

Figure 8. Variables Important to the Projection (VIPs) values for GA selected BCP properties of the common molecular skeleton for the best model (level C) obtained including all MX derivatives.

an opening of the lactone ring when electrophilic MX compounds would react directly with a nucleophilic site on DNA. The prominence of the O_1—C_3 bond as a most likely reactive site also features as a PC at level C as well as at level E. The issue of the favoured role of the α carbon (C_5) over the β carbon (C_4) does not find a totally unambiguous resolution in QTMS. Evidence, shown in Figure 8 and also based on unpublished material, seems to express some preference for the α carbon. For example, the first four highest VIP variables in Figure 8 all involve atom C_3, which is closer to C_5 than to C_4. However, in several models a PC related to the C_2=C_4 bond features high. In summary, QTMS yields good predictivity of the mutagenic activity and produces reasonable evidence in favour of the α carbon, and the lactone ring opening hypothesis.

Conclusions

Two different sets of mutagenic compounds, the triazenes and MX derivatives, are analysed by a new method called quantum topological molecular similarity (QTMS). This method generates bond descriptors from contemporary geometry-optimised *ab initio* wave functions. A feature of QTMS that improves understanding beyond existing QSARs is that it selects the bonds directly involved in the (re)activity. We emphasise that QTMS cannot be reduced to *classical* topological descriptors and should hence be confused with it. A subsequent chemometric analysis, using GA, PC and PLS, delivers statistically scrutinised QSARs. The ability of QTMS to highlight the active center of the mutagens helps in resolving mechanistic ambiguities. The triazene hydroxylation pathway that involves a direct hydrogen abstraction from the methyl is strongly disfavoured. For the MX derivatives more circumstantial evidence points towards a central role in the reactivity of C_α in the C_β=C_α—C=O system, and does not exclude lactone ring opening.

References

1. Voet, D. and Voet, J.G., Biochemistry, 2nd Edition, Wiley, New York, USA, 1995.
2. Maron, D. and Ames, B.N., Mutat. Res., 113 (1983) 173.
3. O'Brien, S.E. and Popelier, P.L.A., J. Chem. Inf. Comp. Sci., 41 (2001) 764.
4. Shusterman, A.J., Debnath, A.K., Hansch, C., Gregory, W.H., Frank, R.F., Greene, A.C. and Watkins, S.F., Mol. Pharm., 36 (1989) 939.
5. Tuppurainen, K. and Lotjonen, S., Mutat. Res., 287 (1993) 235.
6. Bader, R.F.W., Atoms in Molecules. A Quantum Theory. Oxford University Press, Oxford, UK, 1990.
7. Bader, R.F.W., Acc. Chem. Res., 18 (1985) 9.
8. Bader, R.F.W., Chem. Rev., 91 (1991) 893.
9. Popelier, P.L.A., Atoms in Molecules. An Introduction. Pearson Education, London, UK, 2000.

10. Popelier, P.L.A., J. Phys. Chem. A, 103 (1999) 2883.
11. O'Brien, S.E. and Popelier, P.L.A., J. Chem. Soc., Perkin Trans. 2 (2002) 478.
12. Smith, P.J. and Popelier, P.L.A., J. Comput.-Aided Mol. Des., 18 (2004) 135.
13. O'Brien, S.E. and Popelier, P.L.A., ECCOMAS, Barcelona, Spain, 2000.
14. Popelier, P.L.A., Chaudry, U. and Smith, P.J., J. Chem. Soc., Perkin Trans. 2 (2002) 1231.
15. Chaudry, U.A. and Popelier, P.L.A., J. Phys. Chem. A, 107 (2003) 4578.
16. Chaudry, U.A. and Popelier, P.L.A., J. Org. Chem., 69 (2004) 233.
17. Popelier, P.L.A., Chem. Phys. Lett., 228 (1994) 160.
18. Malcolm, N.O.J. and Popelier, P.L.A., J. Comp. Chem., 24 (2002) 437.
19. Bader, R.F.W., Slee, T.S., Cremer, D. and Kraka, E., J. Am. Chem. Soc., 105 (1983) 5061
20. Howard, S.T. and Lamarche, O., J. Phys. Org. Chem., 16 (2003) 133.
21. Cioslowski, J. Mixon, S.T., J. Am. Chem. Soc., 113 (1991) 4142.
22. Bader, R.F.W., Nguyen-Dang, T.T. and Tal, Y., Rep. Prog. Phys., 44 (1981) 893
23. Cheeseman, J.R., Carroll, M.T. and Bader, R.F.W., Chem. Phys. Lett., 143 (1988) 450.
24. Bader, R.F.W. and Preston, H.J.T., Int. J. Quant. Chem., 3 (1969) 327.
25. Schaftenaar, G. and Noordik, J.H., J. Comput.-Aided Mol. Des., 14 (2000) 123.
26. GAUSSIAN98. Gaussian 98, Revision A.7, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Zakrzewski, V. G. Montgomery, J. A., Jr., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J. M., Daniels, A.D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Baboul, A.G., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S. and Pople, J.A., Gaussian, Inc., Pittsburgh, PA, USA, 1998.
27. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
28. Foresman, J.B. and Frisch, A., Exploring Chemistry with Electronic Structure Methods, Gaussian Inc., Pittsburgh, PA USA, 1996.
29. Becke, A.D., J. Chem. Phys., 98 (1993) 5648.
30. 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/.
31. Popelier, P.L.A., Chaudry, U.A. and Smith, P.J., J. Chem. Soc., Perkin Trans. II (2002) 1231.
32. Wold, S., Sjostrom, M. and Eriksson, L., In Schleyer, P., Encycl. of Comp. Chem. Wiley, Chichester, UK, 1998, p. 2006.
33. Wold, S., Kettaneh, N. and Tjessem, K., J. Chemometr., 10 (1996) 463.
34. Holland, J.H., Adaption in Natural and Artificial Systems, MIT Press, Cambridge, MA, USA, 1992.
35. UMETRICS. info@umetrics.com: www.umetrics.com, 2002.
36. Livingstone, D.J., Data Analysis for Chemists, Oxford University Press, Oxford, UK, 1995.
37. Maw, H.H. and Hall, L.H., J. Chem. Inf. Comput. Sci., 41 (2001) 1248.
38. Wise, B.M. and Gallagher, N.B., Eigenvector Research, Manson, WA, USA, 2003.
39. Wold, S., In van de Waterbeemd, H. (Ed.) Chemometric Methods in Molecular Design, VCH, Weinheim, 1995, p. 195.
40. O'Brien, S.E., Dept. of Chemistry, UMIST, Manchester, UK, 2000.
41. SPSS Inc. version 10.0.7 http://www.spss.com: Chicago, IL, USA, 2000.
42. Venger, B.H., Hansch, C., Hatheway, G.J. and Amrein, Y.U., J. Med. Chem., 22 (1979) 473.
43. Shusterman, A.J., Johnson, A.S. and Hansch, C., Int. J. Quant. Chem., 36 (1989) 19.
44. Hansch, C., Drug Metab. Rev., 1 (1972) 1.
45. Pires, J.M., Floriano, W.B. and Gaudio, A.C., J. Mol. Struct., 389 (1997) 159.
46. Fukui, K., Yonezawa, T. and Nagata, C., J. Chem. Phys., 20 (1952) 722.
47. Ho, M., Schmider, H., Edgecombe, K.E. and Smith, V.H., J. Int. J. Quant. Chem., Quant. Chem. Symp., 28 (1994) 215.
48. Popelier, P.L.A., In Ford, M., Livingstone, D.J., Dearden, J. and van de Waterbeemd, H. (Eds.), EuroQSAR2002: Designing Drugs and Crop Protectants: Processes, Problems and Solutions. Blackwell, Oxford, UK, 2003, p. 130.
49. LaLonde, R.T., Cook, G.P., Perakyla, H. and Bu, L., Chem. Res. Toxicol., 4 (1992) 540.
50. Tuppurainen, K., Chemosphere, 38 (1999) 3015.
51. Tupparainen, K., Lotjonen, S., Lattikainen, R., Vartiainen, T., Maran, U., Strandberg, M. and Tamm, T., Mutat. Res., 247 (1991) 97.
52. Tupparainen, K., Lotjonen, S., Laatikainen, R. and Vartiainen, T., Mutat. Res., 266 (1992) 181.
53. Poso, A., Tuppurainen, K. and Gynther, J., J. Mol. Struct.-Theochem., 304 (1994) 255.