Journal of Computer-Aided Molecular Design, 11 (1997) 409–422.

J-CAMD 404

Evaluation of a novel infrared range vibration-based descriptor (EVA) for QSAR studies. 1. General application

David B. Turner[a,*], Peter Willett[a], Allan M. Ferguson[b] and Trevor Heritage[b]

[a]*Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.*
[b]*Shell Research Ltd., Sittingbourne, Kent ME9 8AG, U.K.*

Received 29 January 1997
Accepted 5 April 1997

Keywords: AM1; CoMFA; PM3; IR vibration; Quantitative structure–activity relationships

Summary

A novel molecular descriptor (EVA) based upon calculated infrared range vibrational frequencies is evaluated for use in QSAR studies. The descriptor is invariant to both translation and rotation of the structures concerned. The method was applied to 11 QSAR datasets exhibiting both a range of biological endpoints and various degrees of structural diversity. This study demonstrates that robust QSAR models can be obtained using the EVA descriptor and examines the effect of EVA parameter changes on these models; recommendations are made as to the appropriate choice of parameters. The performance of EVA was found to be comparable in statistical terms to that of CoMFA, despite the fact that EVA does not require the generation of a structural alignment. Models derived using semiempirical (MOPAC AM1 and PM3) and AMBER mechanics calculated normal mode frequencies are compared, with the overall conclusion that the semiempirical methods perform equally well and both outperform the AMBER-based models.

Introduction

Since the advent of the classical QSAR techniques exemplified by the Hansch equation [1], there has been considerable progress in the development of both descriptors and statistical techniques for use in such studies. These include methods which provide explicit descriptions of steric, electronic or hydrophobic properties in 3D space (3D QSAR) [2–8] together with a plethora of quantum-chemical descriptors typically derived using molecular orbital (MO) techniques [9]. Many of these descriptors, at least in principle, permit the development of QSAR models from more heterogeneous datasets than is generally possible with classical QSAR. The classical, 3D and MO-based QSAR techniques can be seen as complementary to one another (see, for example, the molecular shape analysis method [10]), but the more recently developed methods do not at present supplant classical QSAR [11].

Perhaps the most well known of the 3D QSAR techniques is CoMFA [3], which uses the calculated values of steric and electrostatic fields (expressed as interaction energies) sampled at the intersections of a 3D grid placed around the structures concerned. One of the major hurdles to be overcome when applying a method such as CoMFA is that of aligning the structures concerned. In CoMFA a binding hypothesis is developed which is tested through the development of regression models. This process is usually repeated until one or more statistically satisfactory models are obtained. The binding hypotheses require the specification of both a conformation for each structure and the mutual alignment of each structure such that equivalent points (e.g. the putative pharmacophore) are overlaid. One of the problems with this procedure is that very small changes in the relative alignment of the structures involved can result in substantially different regression models [12]. Furthermore, it has been reported that the overall orientation of the structures (taken as a rigid body) within the lattice can also substantially alter modelling statistics [13]. Despite the fact that various techniques have been reported which overcome these difficulties to some extent [12–17], there nonetheless re-

*To whom correspondence should be addressed.

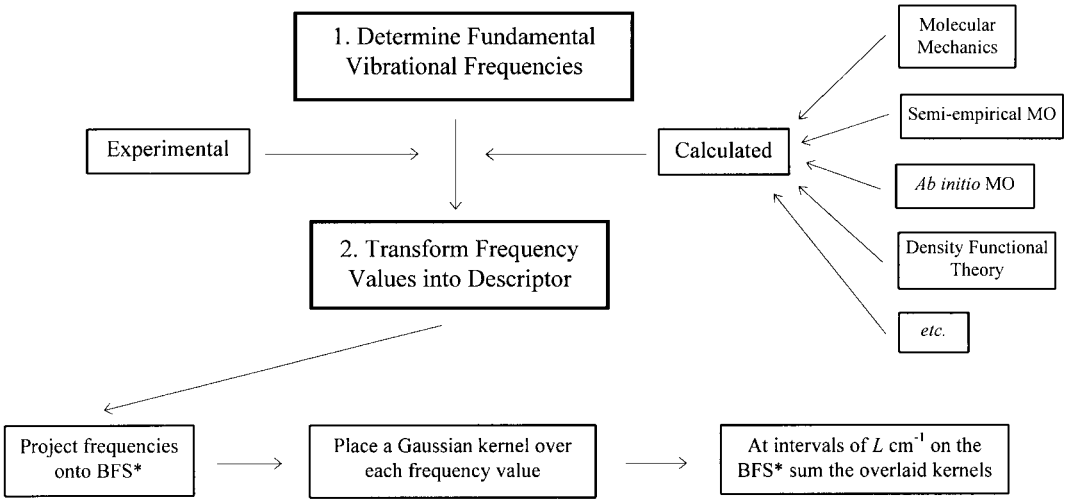

Fig. 1. Overview of the steps required to calculate the EVA descriptor. * BFS 1–4000 cm^{-1}.

main substantial difficulties associated with the use of grids and the need for alignment.

It is clear from the preceding discussion that it is of interest to provide a descriptor which encompasses 3D information but is invariant to considerations of structural overlay and overall orientation. One such approach, based on autocorrelation vectors, has been described by Wagener et al. [18]. Silverman and Platt [16] have recently described an alignment-free method based upon the moments of molecular mass and charge distributions. Alternatively, a method tentatively referred to as 3WD (3D grid WHIM Description) [19] may also provide an effective orientation-independent descriptor. These methods are all based upon alternative descriptions of molecular fields to those used in CoMFA or they describe molecular surfaces only. Here we provide an evaluation of quite a different alignment-free descriptor, referred to as EVA, which is derived from calculated infrared (IR) range vibrational frequencies and which was originally developed by workers at Shell Research Ltd. (SRL) [20,21]. Conceptually, the information content of EVA is intermediate between that of topological and 'true' 3D QSAR descriptors since, whilst conformation is encoded in the descriptor, this information is implicit rather than explicit as it is, for example, with CoMFA. The EVA methodology has previously been described in some detail [20–22] and only a brief summary is given below. An important feature of EVA is that, like CoMFA, the technique can, in principle, be applied to structurally diverse datasets since the models need only be interpolative in terms of the descriptor variables (i.e. field values in CoMFA or derived variables in EVA) for valid predictions to be made of the activity/property of a given compound. Classical QSAR is generally limited by the constraint that for valid predictions to be made for a compound, its structural features must be explicit in the training set; the advantage of this constraint, however, is that a Hansch analysis is more readily interpreted in physicochemical terms than either a CoMFA or an EVA analysis.

Derivation of the EVA descriptor

The main steps involved in the derivation of the EVA descriptor are summarised in Fig. 1. Typically, the normal modes (frequencies and directions of vibration) are calculated using a classical normal coordinate analysis (NCA) [23] of an appropriately energy-minimised structure. The EigenVAlues taken from such an analysis provide the normal mode frequencies from which the EVA descriptor is derived. In the general case a structure will have 3N – 6 eigenvalues (or 3N – 5 for a linear structure such as acetylene, for example), where N is the number of atoms in a molecule. Thus, except in the special case where each structure has the same number of atoms, the number of frequencies will be different for each structure; that is, the property is in nonstandard form. Even where a pair of structures has the same number of vibrations, it is often difficult to establish which vibrations should be compared between the two frequency sets. A novel technique has thus been developed in order to standardise the property such that each compound is characterised by an equivalent-length descriptor. Each frequency for a given structure is first projected onto a bounded frequency scale (BFS) covering a range of 1–4000 cm^{-1}. A Gaussian kernel of fixed standard deviation σ cm^{-1} is then placed over each and every eigenvalue; this results in 3N – 6 (or 3N – 5) overlapping kernels. An appropriate sampling interval (L) is then selected and, starting at 1 cm^{-1}, the BFS is sampled every L cm^{-1}. The value of the EVA descriptor at a sampling point (x) is the summation of the 3N – 6 overlapped Gaussian curves, viz.

$$EVA_x = \sum_{i=1}^{3N-6} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-f_i)^2/2\sigma^2}$$

where f_i is the *i*th frequency for the structure in question. An example of the derivation of the EVA descriptor is shown in Fig. 2, while Fig. 3 illustrates the profile of the descriptor over the whole BFS calculated for the steroid deoxycortisol and using three different σ values. In practice, for efficiency, a heuristic can be used to exclude from consideration those kernels which make a negligible contribution to a given variable; these cutoffs are of course related to the selected σ term. Default values for σ and L have previously been 10 and 5 cm^{-1}, respectively [21], which provides a descriptor consisting of 800 (i.e. 4000/5) variables. With typical QSAR datasets consisting of only 10–100 structures, this is an underdetermined data table and hence there is a requirement to use data reduction techniques such as PLS [24,25] to reduce the dimensionality of the problem; CoMFA, of course, and other multivariate QSAR techniques also use PLS for model development.

The purpose of the EVA standardisation procedure is *not* to simulate an experimental IR spectrum. Transition dipole information has been discarded and, therefore, vibrations that would otherwise be IR-inactive for reasons of symmetry are included in the descriptor. The purpose of the EVA technique is to apply a probability density function to the normal mode frequencies. This function effectively 'smears out' the vibrational frequencies such that vibrations of similar frequency overlap to a greater or lesser extent; this kernel overlap provides the variable variance upon which PLS modelling is based. The use of a fixed σ term means that all normal mode frequencies are equally weighted prior to statistical analysis.

Previous studies with EVA

In-house studies at SRL have indicated that both robust discriminant (active/inactive classifications) and PLS-based regression models can be obtained using the EVA descriptor. These models were developed for a range of proprietary agrochemical datasets and the details of such studies have, therefore, not been published. However, a study using the heterogeneous BC(DEF) dataset of Cramer [26] has been published [21] in which EVA is used to model and predict log P values of the structures concerned. Leave-one-out (LOO) cross-validation statistics using all compounds were encouraging (cross-validated $r^2 = 0.68$) while a subsequent training set/test set division of the structures provided a good test set predictive r^2 of 0.65. This demonstrates that EVA can be used as both an explanatory and a predictive tool with a set of quite diverse compounds.

In the work described herein, the use of EVA QSAR is extended to 11 different datasets, exhibiting a range of biological endpoints, with a view to demonstrating the general potential of the technique; CoMFA models are used as benchmarks against which to judge the statistical performance of EVA. Previous studies [21] have been restricted to quite a narrow range of possible EVA parameters and, therefore, a wide-ranging evaluation of the effect of parameter changes on QSAR modelling results has been performed. A comparative evaluation of the results obtained using normal mode frequencies derived from semiempirical (AM1 and PM3) and AMBER molecular mechanics is also described.

Methods

Notation

The EVA descriptor is described using the following notation: {source, σ, L, S}, where source refers to the means used to optimise the geometry and to calculate the

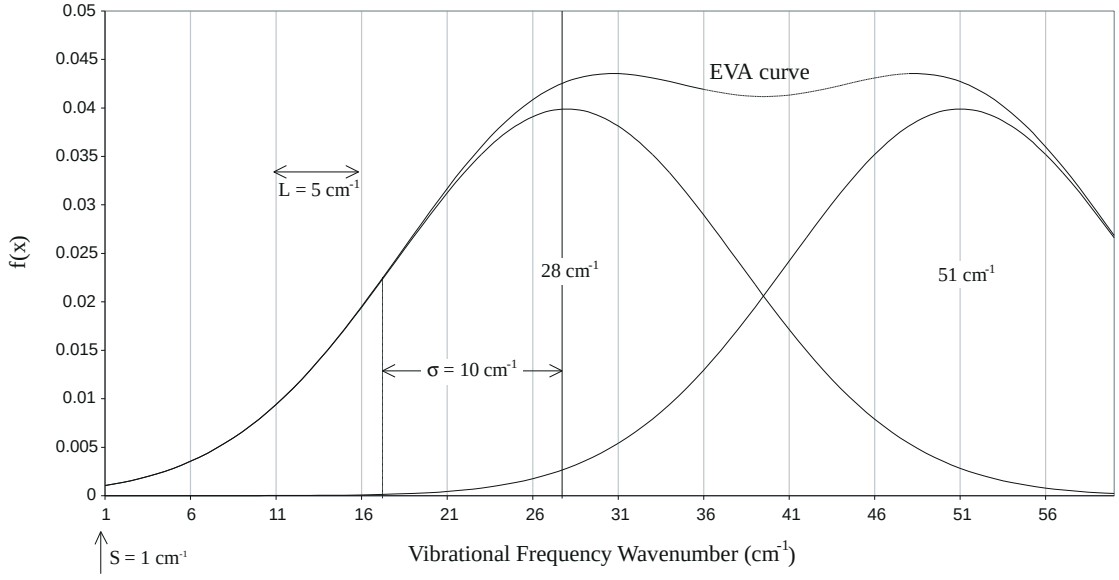

Fig. 2. Details of the calculation of the EVA descriptor for two hypothetical frequencies at 28 and 51 cm^{-1} and using a σ of 10 cm^{-1}; a default sampling increment (L) of 5 cm^{-1} and a reading frame (determined by S) of 1 cm^{-1} have been indicated.

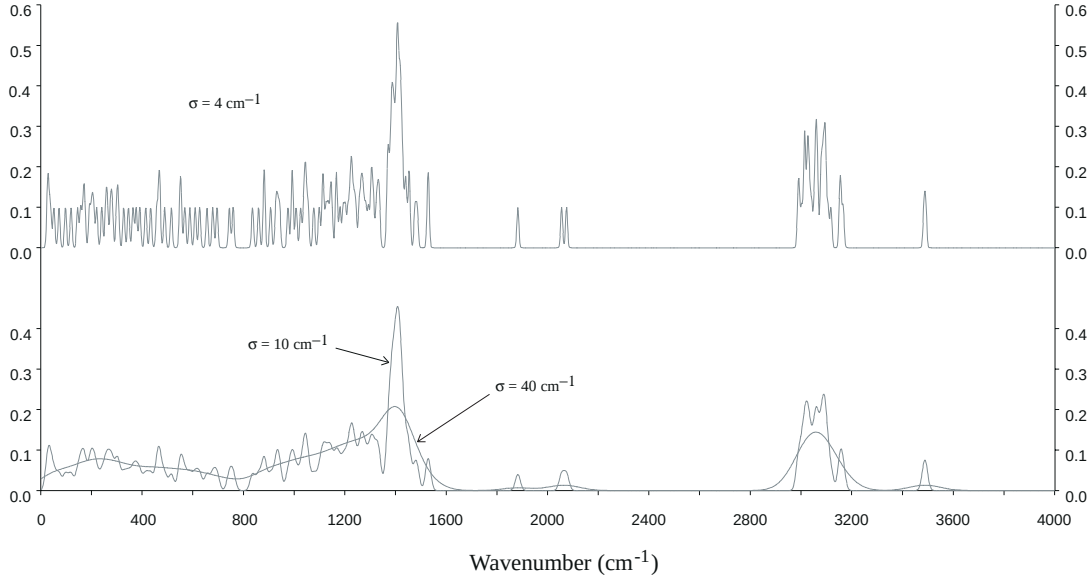

Fig. 3. EVA descriptor profiles (pseudo-spectra) for the steroid deoxycortisol calculated using σ's of 4, 10 and 40 cm^{-1}. Note that the detailed features of the spiky 4 cm^{-1} pseudo-spectrum are gradually smoothed out as σ is raised to 40 cm^{-1}.

normal modes (PM3, AM1, etc.), σ is the Gaussian spread term in cm^{-1}, L is the sampling interval in cm^{-1}, and S is an optional term which refers to the point on the BFS at which sampling is initiated; omission of the S term implies the default value of 1 cm^{-1}.

The following is an alphabetic list of abbreviations related to PLS analyses: F – F-test for significance; LOO – leave-one-out (cross-validation); LV_{opt} – optimum number of latent variables (LVs); q^2 – LOO cross-validated r^2; r^2 – fitted model r^2; SE and SE_{CV} – standard error and cross-validated standard error.

Software and hardware

All the work described herein was carried out using either a Silicon Graphics Iris Indigo R3000 or an Indigo2 R4000. The molecular modelling software used was SYBYL 6.0 [27]. The software required to perform the EVA standardisation process was custom-written in the C programming language.

Datasets

Eleven QSAR datasets (Table 1) were used, taken either from the published literature (for which details are given in the cited references) or from proprietary information provided by SRL. Four of the datasets fall into the latter category and, for reasons of commercial confidentiality, the structures themselves cannot be revealed. The conformations used in CoMFA and as starting points for EVA geometry optimisation were those provided by the original workers except in the case of the cocaines dataset, for which the procedures given in the published study [28] were followed as closely as possible.

The biological activity measurements of the 11 datasets are of two distinct types. In the first category the activities are direct measurements of binding affinity to a receptor, while the second type are endpoints such as those expressed by toxicity indices which tend to correlate with biotransport and distribution properties. Further details on the reproducibility of the target activity data are given in the relevant papers or references therein. The steroid structures [3] are a standard dataset much used in the QSAR literature [18,29–31] and the actual set used, the *Cramer* steroids, is that supplied with the SYBYL modelling software [27]. The steroids have been modelled for two types of binding affinity (Table 1). A second steroid dataset, the *OxMol* steroids [31], which is distributed with the Oxford Molecular [32] software, was also used for some of the work described here. Force-field parameters for the sulphonamide and oxadiazole structures were developed at SRL to supplement those not available in the standard Tripos force field. Missing bond lengths were estimated using a compound for which an X-ray structure was available, while the default SYBYL values were used for bond-stretching force constants (600 kcal/$Å^2$) and torsional terms (0.2 kcal/cos(n, θ), where θ is the torsion angle and n is the periodicity). Approximate force constants are, of course, expected to have an effect upon the absolute values of the vibrational frequencies obtained. However, given that the functional groups to which these approximations are applied are present in each and every structure of the respective datasets, the relative frequency values are expected to be little affected. Therefore, the information content of the resulting EVA descriptor, at least in terms of QSAR analysis, is not anticipated to be significantly altered by these approximations.

As indicated in Table 1, the biological activities of the dibenzofuran, dibenzo-*p*-dioxin and biphenyl datasets are

all for binding to the Ah (dioxin) receptor. As a result, QSAR analyses were performed with each dataset separately and with all the other four possible combinations as has been done previously using CoMFA [33].

Calculation of normal mode frequencies (EVA)

All semiempirical AM1 and PM3 calculations were performed using MOPAC 5.0 [34] via the SYBYL QCPE [35] interface. The conformations used for the CoMFA analyses were adopted as the starting points for the MOPAC geometry optimisation of all structures using either the AM1 or the PM3 Hamiltonian. The MOPAC parameters utilised were default values except where reset through the use of the following MOPAC keywords: full geometry optimisation was performed (ALL_BONDS_AND_ANGLES) with stopping criteria of GNORM = 0.05 and SCFCRT = 10^{-12}. It is noted that the MOPAC semiempirical methods favour the pyramidal form of nitrogen and thus have a tendency to produce such an arrangement in the amide bond of peptides. The experimentally observed form of peptide bonds is planar and a molecular mechanics correction (using the MMOK keyword) is therefore used to favour this configuration. The PARASOK keyword was used for AM1 calculations where sulphur atoms were present because MOPAC 5.0 AM1 lacks parameters for sulphur and thus MNDO parameters are used. There are sulphur atoms in all the sulphonamide structures, one member of the oxadiazole dataset, one of the β-carbolines, one of the piperidines, one of the nitromethylene heterocycles (nmh) and two of the muscarinic structures.

Subsequent to geometry optimisation, a separate MOPAC FORCE calculation was carried out so as to determine the force constants from which the molecular vibrational frequencies are calculated. It is, of course, possible to combine the geometry optimisation and FORCE calculation into a single MOPAC job. However, it was found that, on occasions, this approach resulted in poorer normal mode frequencies (increased numbers of 'negative' frequencies or larger 'negative' values) than when separate geometry and force calculations were performed. The vibrational frequencies were then extracted from the NORMAL COORDINATE ANALYSIS listing in the MOPAC '.out' files. The vibrations were examined to check for any negative ('imaginary') frequencies; these indicate that the structure concerned is not at or sufficiently close to a potential energy surface minimum (a stationary point) and thus further geometry optimisation is required. In some cases, simply repeating the geometry optimisation with the parameters described above resulted in the removal of negative eigenvalues. Otherwise, a more stringent convergence criterion was used (GNORM = 0.01). All structures had no more than one negative 'vibrational frequency' (i.e. eigenvalue) and none had a vibrational frequency of less than -50 cm^{-1}. Negative vibrational frequencies are excluded from consideration when generating the EVA descriptor.

EVA PLS analyses

For each means of geometry optimisation and calculating normal mode frequencies (MOPAC AM1/PM3 and AMBER 3.0 [36]), a systematic evaluation was made of the quality of EVA QSAR models based upon descriptors derived using a wide range of σ values. Descriptors were derived using all values of σ in the set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 16, 18, 20, 22, 24, 26, 28, 30, 40, 50} giving 22 sets of EVA descriptors for each dataset and for each means of calculating the frequencies, resulting in 726 (i.e. $22 \times 3 \times 11$) descriptor sets in total.

The EVA descriptor sets were generated from the MOPAC/AMBER vibrational frequencies and then IMPORTed into a SYBYL molecular spreadsheet together with activity values in order to perform PLS analyses. Full LOO cross-validation was used in all cases to ensure that the results were reproducible. In order to limit computational requirements, a maximum of nine

TABLE 1
DETAILS OF THE 11 QSAR DATASETS

Dataset	n	Biological endpoint	Alignment basis (CoMFA)
Sulphonamides [SRL]	100	log $1/IC_{50}$ for acetolactate synthase inhibition	Steric
Biphenyls [33]	14	Ah (dioxin) receptor binding affinity (pEC_{50})	Steric/electrostatic
Cocaines [28]	13	Cocaine binding site ($1/\log IC_{50}$)	Steric
Dibenzofurans [33]	39	As biphenyls	Steric
β-Carboline [45]	41	Benzodiazepine receptor inverse agonists and antagonists ($\log IC_{50}$)	Steric/pharmacophore
Muscarinics [46]	39	Muscarinic agonists (pD_2)	Steric/pharmacophore
Nitromethylene heterocycles [SRL]	17	$1/\log LC_{50}$ values for the pea aphid	Steric
Oxadiazoles [SRL]	23	Toxicity index (TI) for red spider mite eggs ($1/\log TI$)	Steric
Dibenzo-*p*-dioxins [33]	25	As biphenyls	Steric
Piperidines [SRL]	137	$1/\log IC_{50}$ for *U. Maydis*	Steric
Steroids [3]	21	Testosterone- and corticosterone-binding globulin (TBG and CBG) binding affinity ($-\log [1/K]$)	Steric

SRL: Shell Research Ltd. (proprietary datasets); n: number of compounds in each dataset.

LVs were used for all cross-validation runs, with the exception of some of the largest datasets for which a maximum of seven were used; the SYBYL default is 5. The minimum variance criterion (SYBYL MINIMUM_SIGMA) was set to 0.0 in all cases. The full range of 66 PLS analyses per dataset were carried out with no prior data scaling (SYBYL SCALING_METHOD "NONE"). In addition, all the analyses based on the $\sigma = 10$ cm^{-1} descriptor sets were repeated using pre-autoscaled data (SCALING_METHOD "AUTOSCALE") as were those for all 22 descriptor sets of the biphenyl and oxadiazole datasets. Results are reported as q^2 scores, which are considered to be better indicators of the potential usefulness of the models for the prediction of new compounds than fitted r^2 values since the models are evaluated and selected by the predictions made during cross-validation; PLS model selection criteria are described below. Much more detailed and rigorously validated EVA QSAR analyses (randomisation tests and external test set predictions) are described in the following paper of this series [37]. It is important to acknowledge that, depending on the degree of clustering within a QSAR dataset, the use of LOO cross-validation does not necessarily guarantee selection of the best model [38]. However, the purpose of this work is to demonstrate the potential of EVA as a QSAR descriptor rather than to evaluate different cross-validation techniques, so we have restricted cross-validation to the LOO method.

CoMFA analyses

The CoMFA analyses were all undertaken using standard SYBYL default values, viz. a 2 Å grid spacing in all directions, with the grid extending 4 Å beyond the union molecular volume; a carbon sp^3 probe with a unit positive charge was used; the minimum variance criterion was set to zero; and steric and electrostatic cutoff values were both ±30 kcal/mol, with electrostatic values lying within the steric cutoff region excluded from the analyses. The Gasteiger and Marsili method of calculating charges was utilised. As with the EVA analyses, full LOO cross-validation was used with a maximum of seven or nine LVs. Analyses were done using steric and electrostatic fields, both separately and combined, and were performed for unscaled, pre-autoscaled and pre-blockscaled data as appropriate.

PLS model selection criteria (EVA and CoMFA)

The optimum numbers of LVs (LV_{opt}) indicated by a PLS cross-validation analysis can be selected according to various 'rules-of-thumb' [39]. In SYBYL-PLS the model with the highest F score is reported where F is given by

$$F = \frac{r^2 / A}{(1-r^2)/(M-A-1)} \quad (1)$$

where the r^2 terms can be either the fitted or cross-validated r^2. The F-test score is the ratio of r^2 to $1 - r^2$ (explained to unexplained), weighted so that the fewer the explanatory properties and the more there are of the target values, the higher the F score. However, the formula for calculating cross-validated SE (SE_{CV}) used in SYBYL-CoMFA is

$$SE_{CV} = (PRESS/(M - A - 1)^{1/2} \quad (2)$$

where PRESS is the predictive residual sum of squares, A

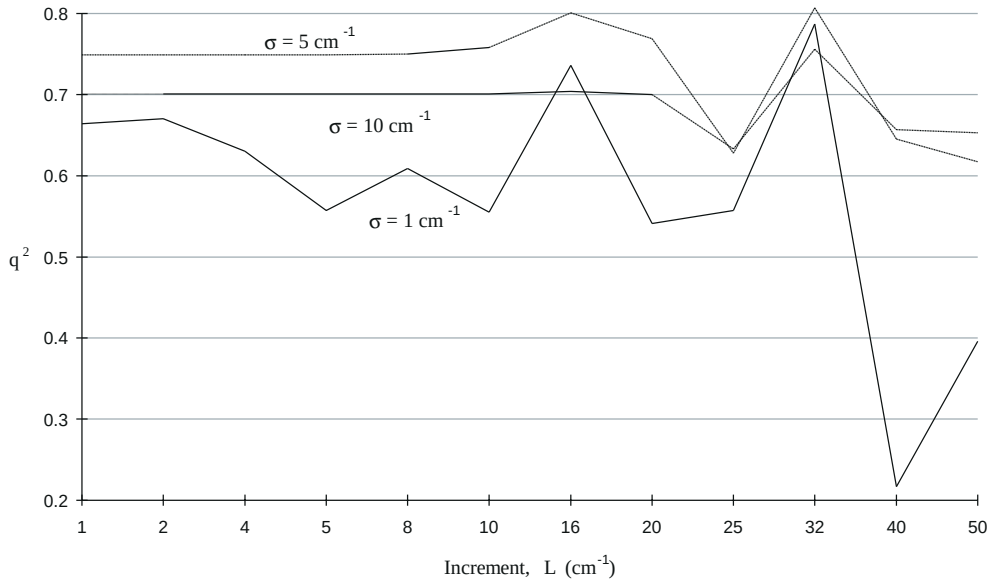

Fig. 4. Plots of the variation of q^2 scores with L. In these analyses σ was fixed and a number of descriptor sets were derived using a range of L values. LOO PLS was then applied and the q^2 scores were extracted. The points at which the q^2 scores become erratic were used to determine the L_{crit} value for each σ; see the main text for a further explanation.

TABLE 2
L_{crit} VALUES FOR SELECTED GAUSSIAN TERMS

Gaussian standard deviation, σ (cm^{-1})	1	2	3	4	6	8	10	14	21
Threshold increment, L_{crit} (cm^{-1})	2	4	5	8	10	16	20	25	32

In order to avoid sampling errors, the value of L should be chosen to be less than L_{crit} for a given σ. These values have been chosen such that effects resulting from the choice of sampling frame (determined by the reading frame, S) are accounted for.

is the number of PLS LVs and M is the number of compounds in the dataset. This penalises models with the same PRESS but more LVs, and means that the q^2 maximum and SE_{CV} minimum do not always occur at the same number of LVs. Wold et al. [24] suggest that, in the interests of model parsimony, LV_{opt} should be selected on the basis of the first SE_{CV} minimum. The rationale behind this is that the more LVs used to describe the training data then, despite the very high conventional r^2 scores often obtained, the greater the danger of overfit to the data. The term overfit is used to mean that there is an expectation of poor predictive performance on a separate test dataset; that is, the model may have become specific for the training set rather than generalised to cope with an external test set. An alternative route to model parsimony has been suggested by Kubinyi and Abraham [39] in which the model with fewer LVs should be selected whenever the increase in q^2 score is less than 5%. There are, however, no fixed rules for determining LV_{opt} and, as demonstrated by, for example, Turner et al. [37], parsimonious model selection does not always result in the best performance on a separate test set. The models reported here are based upon the SYBYL SE_{CV} minimum subject to the conditions given below.

Wold et al. [24] also indicate a further constraint on the value of LV_{opt} insomuch that the number of LVs used to describe the data should not exceed M/4, where M is the number of examples (compounds) in the training set. The use of a compound/LVs ratio of greater than M/4 results in an unacceptably high risk of chance correlation, i.e. an underdetermined data table [40]. Therefore, this constraint has been used as an upper bound when selecting LV_{opt}; the number of compounds in each dataset is indicated in Table 1.

Finally, it is not acceptable to make predictions of the biological activity of structures to greater precision than the error in (reproducibility of) the original measurements. This factor has not been directly addressed in these evaluations. However, the original publications (or SRL studies) deemed the CoMFA predictions to be acceptable and thus the EVA predictions should be treated with caution (in this sense) only if they significantly statistically outperform the CoMFA models.

Results and Discussion

Initially we focus upon the QSAR modelling results obtained using EVA descriptors based upon a Gaussian σ term of 10 cm^{-1}; this has previously been taken as a default value for the σ term [21]. Subsequently, we describe the results of the systematic variation of σ in terms of the effect upon QSAR model statistics. CoMFA analyses were also performed so as to provide benchmarks for the EVA QSAR results. Finally, the results are given of the comparison of EVA QSAR models obtained using AM1-, PM3- and AMBER-derived frequencies.

Selection of appropriate EVA standardisation parameters

There are three parameters to be selected when deriving the EVA descriptor from a set of normal mode frequencies. The most important of these is the EVA σ term, which determines the extent of inter- and intrastructural kernel overlap; this is discussed below. Having selected a σ term to use, there are two further considerations, viz. the BFS sampling increment, L, and the reading frame, S; the latter is determined by the point on the BFS at which sampling of the overlapped kernels is initiated (Fig. 2). The values of the L and S parameters determine the points on the BFS at which sampling of the kernels takes place.

Sampling increment (L) and reading frame (S)

It is intuitive that the sampling increment should be sufficiently small with respect to the selected EVA σ term such that there is no information loss; i.e. each Gaussian kernel should be sampled at a point on the BFS where it has significant value (the way in which significance is assessed is described below). Information loss in this manner amounts to variable pre-selection and should thus be avoided since the variables concerned may be signal, that is have a consistent relationship with the target biological activity. This leads to the concept of critical values of L (L_{crit}) which are σ-specific and which, if exceeded, mean that a sampling error is introduced. At the same time, L should be kept as large as possible so as to minimise computational and storage requirements which may be important where a small σ term (and hence L value) has been used or where a very large dataset is to be modelled. Maximising L will also help to minimise the number of correlated variables as far as possible.

In order to determine values for L_{crit}, the σ parameter was fixed and a number of EVA descriptor sets were derived by systematically increasing L in small steps. LOO cross-validation was then performed with each descriptor set. The L_{crit} values were determined from the points at which the resultant QSAR model statistics

began to show erratic behaviour; that is, fluctuations in the LOO q^2 scores indicating the different information content of the EVA descriptor sets (as judged by PLS in combination with LOO cross-validation). This procedure was repeated for descriptors derived using a wide range of σ terms. Figure 4 shows the variation of q^2 scores with L for a selection of three σ terms using the OxMol steroid dataset and CBG binding activity. The L_{crit} value for a σ term of 10 cm^{-1} is thus about 20 cm^{-1}, while that for the 5 cm^{-1} σ term is around 10 cm^{-1}, and that for the 1 cm^{-1} σ term is about 2 cm^{-1}. In the latter case there are dramatic changes in the q^2 scores as L is increased above 2 cm^{-1}, reflecting the arbitrary omission/inclusion of signal that can occur when $L \gg \sigma$. On the other hand, the q^2 scores are extremely consistent where $L < L_{crit}$. In order to ensure that these L_{crit} values are not specific to the default reading frame (determined by S), a further set of tests was done in which the value of S was systematically altered relative to the default of 1 cm^{-1}. Thus, for a fixed L of, for example, 5 cm^{-1}, S was incremented from 1 to 5.5 cm^{-1} in 0.5 cm^{-1} steps, resulting in nine descriptor sets, in this instance, to which LOO cross-validated PLS was applied as previously. This procedure was again applied over a wide range of σ terms, and L and S values; the results are summarised in Table 2, where the L_{crit} values for various σ terms are given. These results confirm that the intuitively reasonable selection of an L of 5 cm^{-1} with a σ term of 10 cm^{-1} is an acceptable parameter combination. In fact, the results indicate that any value of L up to 20 cm^{-1} can be used with a σ term of 10 cm^{-1} without apparent information loss (change in q^2 values). All subsequent models reported are those for which the relevant L_{crit} is not exceeded and for which S is the default 1 cm^{-1}.

The establishment of these L_{crit} values is important not least because one of the problems with CoMFA at present is that the coarse grid-spacing (typically, 2 Å) that is generally used is such that there is incomplete sampling of molecular fields (information loss). The consequence of this is that the reorientation of an aligned set of compounds as a rigid body within the defining CoMFA 3D region very often results in substantial changes to QSAR modelling performance (as evidenced in the q^2 values, for

TABLE 3
SUMMARY OF EVA QSAR ANALYSES USING A DEFAULT σ OF 10 cm^{-1}

Dataset	Pre-scaling	Source of vibrational frequencies {10, 5, 1}											
		AMBER				MOPAC AM1				MOPAC PM3			
		q^2	r^2	SE	F	q^2	r^2	SE	F	q^2	r^2	SE	F
β-Carbolines[a]	None	0.29 (6)	0.97	0.50	180.6	0.50 (6)	0.97	0.57	195.5	0.39 (3)	0.84	1.30	66.2
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Biphenyls	None	0.16 (1)	0.72	0.48	30.3	≤0	–	–	–	≤0	–	–	–
	Auto	≤0	–	–	–	0.28 (2)	0.90	0.30	49.0	0.28 (1)	0.75	0.46	35.8
Cocaines	None	0.57 (2)	0.91	0.26	51.9	0.49 (2)	0.95	0.20	94.0	0.57 (2)	0.94	0.22	75.8
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Dibenzo-*p*-dioxins	None	0.48 (2)	0.85	0.59	61.5	0.65 (2)	0.85	0.59	60.2	0.65 (2)	0.88	0.53	76.9
	Auto	≤0	–	–	–	0.68 (2)	0.88	0.53	77.2	0.76 (2)	0.91	0.44	115.9
Dibenzofurans	None	0.61 (1)	0.74	0.69	103.4	0.73 (4)	0.96	0.28	201.1	0.61 (2)	0.80	0.62	71.1
	Auto	≤0	–	–	–	0.78 (4)	0.97	0.25	273.9	0.67 (2)	0.83	0.57	86.7
Muscarinics	None	0.42 (3)	0.88	0.27	81.7	0.53 (4)	0.95	0.17	171.3	0.35 (2)	0.81	0.34	74.1
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Nmh	None	0.47 (2)	0.86	0.59	41.3	0.49 (3)	0.93	0.41	61.5	0.54 (3)	0.96	0.30	117.2
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Oxadiazoles	None	≤0	–	–	–	≤0	–	–	–	0.20 (3)	0.96	0.09	141.8
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Piperidines	None	0.71 (5)	0.84	0.42	137.9	0.76 (4)	0.84	0.43	174.7	0.75 (4)	0.84	0.42	177.8
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Steroids	None	0.42 (5)	0.99	0.17	206.4	0.70 (4)	0.98	0.20	175.0	0.32 (2)	0.84	0.51	46.4
(TBG[b] activity)	Auto	0.36 (4)	0.98	0.21	163.8	0.66 (4)	0.98	0.20	177.0	0.15 (1)	0.65	0.74	35.1
Steroids	None	0.79 (2)	0.90	0.38	85.0	0.70 (2)	0.87	0.45	59.1	0.67 (3)	0.94	0.32	84.3
(CBG[c] activity)	Auto	0.77 (2)	0.91	0.37	90.0	0.76 (1)	0.84	0.48	100.9	0.68 (2)	0.91	0.38	86.0
Sulphonamides[d]	None	–	–	–	–	0.54 (6)	0.80	16.4	60.2	0.51 (6)	0.80	16.3	61.3
	Auto	–	–	–	–	0.44 (2)	0.66	20.6	94.8	≤0	–	–	–

Both results unscaled and autoscaled data are used. The LOO q^2 values are reported together with the optimal with the optimal number of LVs in parentheses. All q^2 values of ≤0 are indicated as ≤0 and LV_{opt} is omitted as meaningless. Models are based on the selection of LV_{opt} by minimum SE_{CV} score. Full (fitted) models were derived only where $q^2 > 0$.

[a] AMBER had the required force-field parameters for only 39 structures.

[b] Testosterone-binding globulin affinity as the target activity.

[c] Corticosterone-binding globulin affinity as the target activity.

[d] AMBER force-field parameters not available.

TABLE 4
SUMMARY OF CoMFA QSAR ANALYSES

Dataset	Pre-scaling	Both fields				Steric field only				Electrostatic only			
		q^2	r^2	SE	F	q^2	r^2	SE	F	q^2	r^2	SE	F
β-Carbolines	None	0.68 (4)	0.89	1.12	69.9	0.63 (5)	0.90	1.03	64.4	0.46 (4)	0.78	1.55	32.3
	Block	0.62 (3)	0.87	1.20	80.6	–	–	–	–	–	–	–	–
	Auto	0.56 (2)	0.71	1.73	47.0	0.25 (1)	0.47	2.31	35.0	0.33 (2)	0.50	2.29	18.8
Biphenyls	None	0.49 (3)	0.87	0.36	21.9	0.47 (3)	0.86	0.37	20.7	0.43 (2)	0.78	0.44	19.7
	Block	0.57 (2)	0.83	0.39	27.0	–	–	–	–	–	–	–	–
	Auto	0.47 (3)	0.87	0.36	21.6	0.27 (3)	0.88	0.35	23.4	0.42 (2)	0.80	0.43	21.7
Cocaines	None	0.59 (4)	0.88	0.28	63.0	0.57 (3)	0.84	0.32	62.7	0.57 (2)	0.71	0.43	44.1
	Block	0.56 (2)	0.74	0.40	52.3	–	–	–	–	–	–	–	–
	Auto	0.51 (2)	0.67	0.46	35.9	0.42 (2)	0.68	0.45	37.4	0.54 (3)	0.71	0.43	29.1
Dibenzo-*p*-dioxins	None	0.66 (1)	0.80	0.66	92.3	0.65 (1)	0.80	0.67	89.5	0.36 (2)	0.75	0.75	33.4
	Block	0.73 (2)	0.88	0.52	80.6	–	–	–	–	–	–	–	–
	Auto	0.39 (1)	0.73	0.77	60.7	0.36 (1)	0.68	0.84	48.7	0.32 (2)	0.67	0.87	22.1
Dibenzofurans	None	0.72 (6)	0.85	0.57	29.9	0.63 (3)	0.77	0.67	38.8	0.73 (4)	0.85	0.56	46.4
	Block	0.73 (3)	0.84	0.56	59.2	–	–	–	–	–	–	–	–
	Auto	0.72 (3)	0.83	0.57	56.8	0.70 (3)	0.82	0.59	52.8	0.77 (3)	0.85	0.55	63.7
Muscarinics	None	0.59 (4)	0.84	0.31	46.0	0.61 (4)	0.85	0.31	47.6	0.52 (3)	0.67	0.44	23.8
	Block	0.59 (4)	0.84	0.31	45.4	–	–	–	–	–	–	–	–
	Auto	0.51 (4)	0.76	0.39	26.2	0.51 (2)	0.68	0.43	38.0	≤0	–	–	–
Nmh	None	0.84 (3)	0.96	0.34	92.4	0.87 (3)	0.96	0.32	105.3	≤0	–	–	–
	Block	0.82 (5)	0.98	0.24	118.3	–	–	–	–	–	–	–	–
	Auto	≤0	–	–	–	≤0	–	–	–	≤0	–	–	–
Oxadiazoles	None	0.51 (2)	0.85	0.07	56.2	0.51 (2)	0.84	0.07	52.8	≤0	–	–	–
	Block	0.39 (2)	0.83	0.08	48.0	–	–	–	–	–	–	–	–
	Auto	0.08 (2)	0.75	0.09	29.6	≤0	–	–	–	0.04 (1)	0.38	0.14	12.7
Piperidines	None	0.73 (3)	0.80	0.48	175.2	0.73 (3)	0.80	0.48	174.1	0.57 (3)	0.71	0.57	110.3
	Block	0.73 (3)	0.80	0.48	176.1	–	–	–	–	–	–	–	–
	Auto	0.55 (2)	0.67	0.61	134.4	0.51 (2)	0.68	0.60	145.0	0.33 (3)	0.53	0.73	49.9
Steroids	None	0.62 (3)	0.92	0.37	67.7	0.59 (3)	0.87	0.47	39.3	0.65 (2)	0.90	0.40	80.6
(TBG activity)	Block	0.68 (5)	0.99	0.16	233.9	–	–	–	–	–	–	–	–
	Auto	0.71 (4)	0.96	0.27	95.7	0.28 (1)	0.49	0.89	18.2	0.61 (5)	0.96	0.27	75.4
Steroids	None	0.75 (2)	0.91	0.37	93.0	0.72 (2)	0.89	0.40	75.4	0.77 (4)	0.96	0.25	105.0
(CBG activity)	Block	0.59 (1)	0.70	0.66	44.0	–	–	–	–	–	–	–	–
	Auto	0.75 (3)	0.95	0.29	106.1	0.52 (1)	0.64	0.72	34.3	0.40 (1)	0.61	0.77	29.1
Sulphonamides	None	0.63 (5)	0.82	15.1	87.7	0.63 (5)	0.82	15.3	86.1	0.50 (6)	0.77	17.4	51.6
	Block	0.63 (5)	0.83	15.1	89.0	–	–	–	–	–	–	–	–
	Auto	0.41 (3)	0.64	21.5	56.0	0.31 (2)	0.51	24.7	51.3	0.37 (2)	0.49	25.4	46.1

See Table 3 for further information.

example). This problem has been encountered in our studies [37] while Cho and Tropsha [13] have presented a means of dealing with this feature. EVA does not suffer from sampling-related errors, provided the tabulated, σ-specific L_{crit} values are not exceeded.

Results with EVA default parameters

In this section a brief description is given of the results obtained with the default EVA parameters ($\sigma = 10\ cm^{-1}$); this provides a basis upon which to evaluate the effects of the use of alternative σ terms. Table 3 is a summary of the EVA PLS results, while those for analogous CoMFA analyses are given in Table 4. It is clear that EVA provides reasonably good models (in terms of q^2 scores) for the majority of the datasets. The exceptions to this are the oxadiazole and the biphenyl datasets, for which either a poor or no model is obtained. The CoMFA model scores are, of course, all quite good since this was the basis upon which the datasets were selected. It is interesting to note that, like EVA, the poorest CoMFA q^2 scores are obtained with the oxadiazole and biphenyl datasets (combined field). In most cases the CoMFA electrostatic-field-only models are poorer than those where the steric field is included, thus indicating the different information content of the two field types; this may be a 'true' difference or it may be a consequence of the method used to calculate charges. This finding probably reflects the fact that the datasets (with the exception of the biphenyls) were aligned on a steric basis (Table 1). In most cases the effect of pre-autoscaling the descriptor is to reduce q^2 scores to ≤0, with the notable exceptions of the dibenzofuran, dibenzo-*p*-dioxin, biphenyl and steroid (CBG and TBG activity) datasets. In the case of the biphenyls, and AM1- or PM3-derived descriptors, autoscaling improves

the model from no model at all ($q^2 < 0$) to a model that is poor but (arguably) significant ($q^2 = 0.28$). It should be pointed out that, as discussed below, with the biphenyls much better AM1- and PM3-derived q^2 scores (≈0.45) are obtained for σ values other than 10 cm^{-1} (Table 5). Indeed, for nearly all the datasets there are instances where higher q^2 scores are obtained using σ values other than the default of 10 cm^{-1}; this is discussed below. It is also interesting to note the large differences that in some cases occur between models based on AM1- and PM3-derived frequencies; this is also discussed below.

Finally, in this section, Table 6 summarises the results of combining the three datasets for which the biological activity measurement is that of binding to the Ah (dioxin) receptor – the biphenyl, dibenzofuran and dibenzo-*p*-dioxin compounds. For each of the four possible combinations of datasets, the LOO q^2 results are entirely comparable with those of analogous CoMFA analyses, albeit depending to some extent on the pre-scaling option selected. These results indicate that, like CoMFA, EVA can be effective where heterogeneous sets of structures are combined into one analysis.

Effect of the EVA σ term on QSAR model statistics

As stated previously, the fundamental parameter involved in deriving the EVA descriptor from normal mode frequencies is the σ term. The effect of the use of various σ terms is illustrated in Fig. 3, where it can be seen that the features of the descriptor profile are progressively smoothed as the σ term is raised. The value of σ (the Gaussian standard deviation) determines the extent of both the inter- and intrastructural Gaussian kernel overlap. In this sense, kernels are considered to overlap only if they do so at nonnegligible values. Interstructural overlap enables frequencies of similar value to contribute significantly to a given EVA variable. Intrastructural kernel overlap can generally be seen to be undesirable since this means that more than one normal mode frequency is significantly contributing to the EVA variable(s) concerned and that there is, therefore, a mixing of information. However, in order to provide sufficient interstructural kernel overlap, it is inevitable that there is a certain amount of significant intrastructural overlap. In this section, and on the basis of the results obtained with default EVA parameters, a systematic study is made using a wide range of σ terms so as to determine whether there are significant differences in the resulting QSAR models. Extensive listings of the results are described by Turner [41]; Table 5 provides a summary of the main findings, where σ is extended up to 40 cm^{-1}. For two of the datasets, σ was further extended to 155 cm^{-1} (cocaines) and 230 cm^{-1} (OxMol steroids).

A close examination of Table 5 indicates that for a number of the datasets, an improvement in q^2 scores can

TABLE 5
SUMMARY OF THE BEST-PERFORMING EVA QSAR ANALYSES

Dataset	Pre-scaling	AMBER			MOPAC AM1			MOPAC PM3		
		Best		σ = 10 cm^{-1}	Best		σ = 10 cm^{-1}	Best		σ = 10 cm^{-1}
		q^2	σ	q^2	q^2	σ	q^2	q^2	σ	q^2
β-Carbolines	None	–	–	–	0.66 (7)	22	0.50 (6)	0.39 (2)	2	0.39 (3)
								0.40 (4)	13	
Biphenyls	None	0.18 (1)	13	0.16 (1)	0.14 (3)	7	≤0	0.47 (3)	5	≤0
	Auto	–	–	–	0.45 (3)	16	0.28 (2)	–	–	–
Cocaines	None	0.58 (2)	8	0.57 (2)	0.68 (3)	65 (+)	0.49 (2)	0.60 (2)	6, 7	0.57 (2)
		0.63 (3)	50 (+)		0.55 (3)	13				
Dibenzo-*p*-dioxins	None	0.64 (3)	4	0.53 (3)	0.70 (2)	18–40	0.65 (2)	0.72 (2)	24–26	0.65 (2)
		0.68 (4)	24–26							
Dibenzofurans	None	0.62 (1)	6–8	0.63 (2)	0.74 (4)	7–9	0.73 (4)	0.70 (2)	3	0.61 (2)
		0.63 (2)	10–13							
Muscarinics	None	0.76 (5)	40 (+)	0.42 (3)	0.53 (4)	10	As best	0.39 (4)	40 (+)	0.35 (2)
Nmh	None	0.58 (3)	5	0.47 (2)	0.66 (3)	4	0.49 (3)	0.57 (3)	6–7	0.54 (3)
Oxadiazoles	None	≤0	–	≤0	≤0	–	≤0	0.23 (3)	7–8	0.20 (3)
	Auto	–	–	≤0	≤0	–	≤0	–	–	≤0
Piperidines[a]	None	–		0.71 (5)	0.78 (3)	2–4	0.76 (4)	0.74 (3)	2–3	0.77 (7)
								0.77 (7)	9–13	
Steroids (TBG)	None	0.43 (5)	11	0.42 (5)	0.70 (4)	8–11	0.70 (4)	0.67 (5)	24–28	0.32 (2)
Steroids (CBG)	None	0.83 (2)	4–5	0.79 (2)	0.75 (1)	3	0.70 (2)	0.71 (2)	4–5	0.67 (3)
Sulphonamides	None	–	–	–	0.55 (3)	2	0.56 (7)	0.53 (4)	6–7	0.49 (4)
					0.57 (7)	7–8				

For each dataset this table gives the highest q^2 score and the Gaussian standard deviation at which this score was obtained; the q^2 score for the default Gaussian standard deviation (10 cm^{-1}) is also given. The optimal number of LVs are in parentheses.
See footnotes a–d in Table 3.
[a] AMBER analysis was performed using EVA {10, 5}-based descriptors only.

TABLE 6
SUMMARY OF EVA AND CoMFA QSAR ANALYSES WITH THE DIOXIN (Ah) RECEPTOR-BINDING STRUCTURES

Dataset	Pre-scaling	EVA {AM1, 10, 5}				CoMFA (both fields)			
		q^2	r^2	SE	F	q^2	r^2	SE	F
Dioxins and furans	None	0.63 (3)	0.84	0.60	103.7	0.80 (7)	0.90	0.49	73.4
	Block	–	–	–	–	0.78 (6)	0.90	0.51	79.1
	Auto	0.71 (4)	0.90	0.48	135.1	0.60 (3)	0.73	0.78	54.1
Dioxins and biphenyls	None	0.71 (3)	0.92	0.47	132.6	0.72 (3)	0.86	0.63	69.3
	Block	–	–	–	–	0.76 (5)	0.94	0.44	96.0
	Auto	0.73 (3)	0.93	0.44	155.0	0.82 (4)	0.92	0.48	99.1
Furans and biphenyls	None	0.66 (2)	0.81	0.57	108.7	0.71 (7)	0.89	0.46	53.1
	Block	–	–	–	–	0.75 (4)	0.87	0.49	77.8
	Auto	0.70 (2)	0.86	0.49	152.9	0.55 (3)	0.79	0.61	62.4
Dioxins, furans and biphenyls	None	0.62 (3)	0.83	0.61	123.5	0.62 (3)	0.75	0.75	74.6
	Block	–	–	–	–	0.71 (6)	0.88	0.54	82.9
	Auto	0.68 (3)	0.87	0.55	158.0	0.58 (3)	0.72	0.79	63.9

See Table 3 for further information.

be obtained through the use of descriptors other than those based on $\sigma = 10$ cm^{-1}. For the AM1-derived models this applies to the cocaines, biphenyls, β-carbolines and nitromethylene heterocycles, where differences of 0.16–0.17 q^2 units are observed. For the PM3-derived models the biphenyls provide the greatest difference between the best and default models, where the improvement is from worse than no model at all to a q^2 score of 0.47. In addition, for the modelling of steroid TBG activity with PM3-based modes, there is an improvement of 0.35 in the q^2 score if σ terms of around 26 cm^{-1} are used rather than 10 cm^{-1}. Turning to the AMBER-based models, the greatest difference occurs for the muscarinic dataset, where a model based on a σ of 40 cm^{-1} gives a q^2 score of 0.76 as opposed to 0.42 with a σ of 10 cm^{-1}. There is also an enhancement in q^2 scores of 0.11–0.15 with the dibenzo-*p*-dioxin and nmh datasets. Where σ is extended beyond 40 cm^{-1} (OxMol steroid and cocaine datasets), it appears that PLS remains capable of deriving adequate models even at extremely large σ's. With the OxMol steroids (Fig. 5) it appears that the simplest model is given at a σ of about 4 cm^{-1} (one latent variable q^2 score of 0.72), while at higher σ's this level of score is only retained through the addition of successively more LVs. This trend is, however, not evident with the cocaine dataset (Fig. 6) or with most of the other datasets used herein (not illustrated).

Overall, it seems that in most cases there are only small gains to be had by using σ terms other than 10 cm^{-1}, although there are a number of notable exceptions as indicated above. In general, therefore, it would seem reasonable to start a modelling study with EVA descriptors based on $\sigma = 10$ cm^{-1} and, if modelling is not satisfactory, supplement this with analyses based on σ terms of, say, at about 5, 25 and 50 cm^{-1}. The use of Gaussian spread terms greater than about 50 cm^{-1} has been found

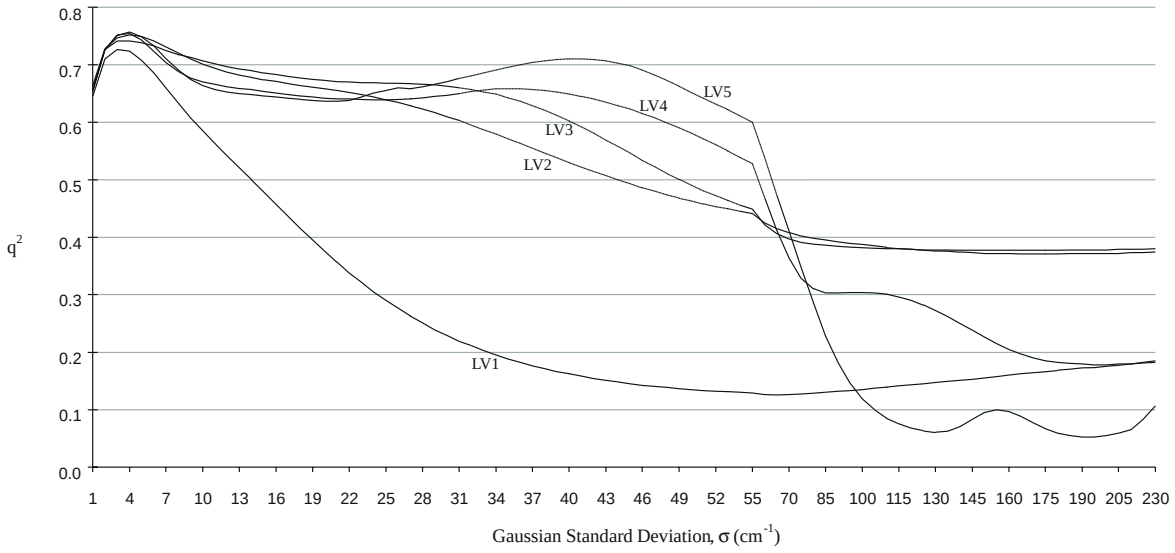

Fig. 5. Cumulative fraction of bioactivity variance (q^2) explained by successive PLS LVs as a result of LOO cross-validation versus Gaussian standard deviation, using sampling increments below the relevant L_{crit} (OxMol steroid dataset (AM1-derived normal modes)).

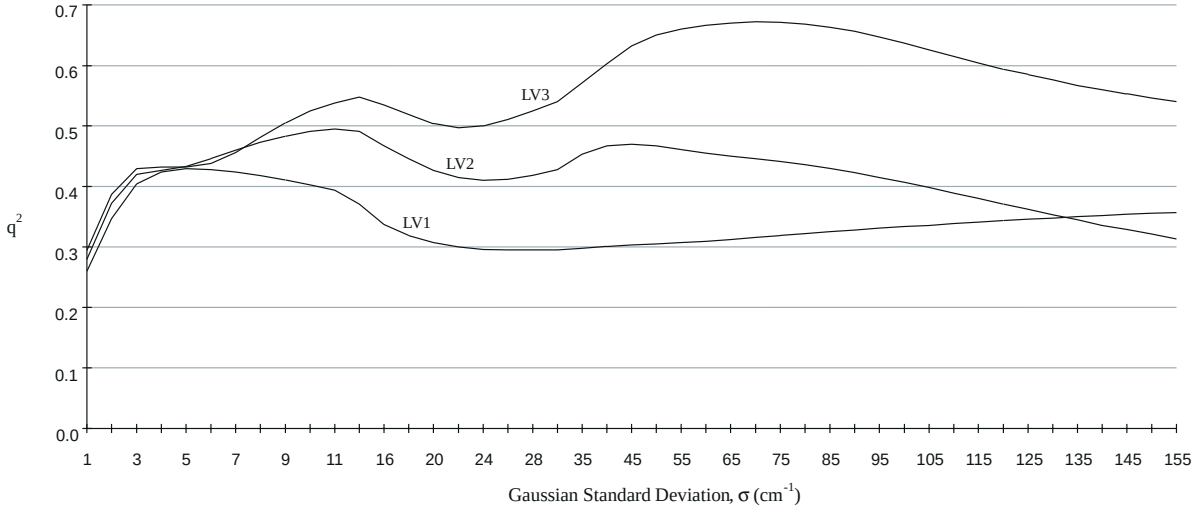

Fig. 6. Cumulative fraction of bioactivity variance (q^2) explained by successive PLS LVs as a result of LOO cross-validation versus Gaussian standard deviation, using sampling increments below the relevant L_{crit} (cocaine dataset (AM1-derived normal modes)).

to be effective at producing high q^2 scores in one case (the cocaines, with a peak at $\sigma = 65$ cm^{-1}). However, with such large σ values it is difficult to envisage the information encoded in the descriptor. Furthermore, the interpretation of such models (backtracking to (contra-)indicated structural features) must be inherently more complex than those based upon small σ's, for which there will tend to be much more of a one-to-one correspondence between an EVA variable and an underlying normal mode frequency.

Oxadiazole dataset

Of the 11 datasets, the oxadiazole dataset is the only case where a significant EVA QSAR model could not be obtained. This dataset is also that for which the CoMFA model is one of the poorest ($q^2 = 0.51$) and the observed correlation is entirely due to the steric field (Table 4). A randomised permutation test [20,42] suggests high confidence in the CoMFA model (i.e. there is a very low estimated probability that chance correlation has occurred) and an examination of the original source of activity data indicates high confidence in (high reproducibility of) these values. In an alternative vein, MOPAC 5.0 is not parameterised for sulphur and we have thus used MNDO parameters to model this structure (MOPAC PARASOK keyword). There is a single sulphur-containing compound in this dataset and there is the possibility that the use of MNDO parameters may have produced peculiar results. However, the predictive residual for this compound from LOO cross-validation is not particularly poor and omission of this compound completely makes little difference to the LOO q^2 score. Furthermore, an examination of the residuals for all compounds indicates that neither a single compound nor a small group of compounds can be said to be outliers that are exerting great leverage on the EVA cross-validation results. Thus, the reason for the lack of any reasonable QSAR model with the oxadiazole dataset is at present not clear to us.

Comparison of models based on AM1-, PM3- or AMBER-calculated frequencies

AM1 and PM3 For a number of datasets there are quite large differences between models based on AM1- and PM3-derived frequencies. The most obvious of these is with the Cramer steroid dataset for TBG binding activity, where, for unscaled data, the difference is nearly 0.40 q^2 units when $\sigma = 10$ cm^{-1} (Table 3), although this difference is considerably reduced if the best-performing models are compared (Table 5). There are also notable differences (>0.1 q^2 units) between the β-carboline, dibenzofuran, muscarinic and oxadiazole AM1- and PM3-based models. These findings indicate that the QSAR modelling results are in some cases dependent upon the means used to optimise geometries and calculate normal modes. However, these differences can to some extent be minimised through the appropriate choice of Gaussian σ.

AMBER and semiempirical methods The performance of AMBER-based models is equivalent to that of AM1-based models for the biphenyl and oxadiazole datasets, for both of which the correlations for either method are anyway very poor. AMBER outperforms AM1 with the muscarinic dataset and to a lesser extent with the steroids for CBG binding activity. Otherwise, AM1 outperforms AMBER with the remainder of the datasets, albeit by a small margin (with the exception of the steroids for TBG binding activity).

A similar picture is indicated in a comparison of AMBER and PM3 results, with the muscarinic dataset again modelled better by AMBER while PM3 performs significantly better with the biphenyl, oxadiazole and steroid (TBG binding activity) datasets. With the cocaines and nitromethylene heterocycles the modelling perform-

ance is almost identical, whilst with the remaining three datasets PM3 performs marginally better than AMBER.

Overall, the semiempirical methods each individually outperform AMBER in terms of cross-validation statistics of the derived QSAR models, with the caveat that AMBER-based descriptors give better results in a few cases. Perhaps, the most important observation is that different means of theoretical modelling (force constant determination) can, on occasions, result in substantially different QSAR models.

Comparison of EVA with CoMFA

In overall terms the statistical performance of the EVA QSAR models is remarkably similar to that of CoMFA in terms of q^2 scores, particularly for the semiempirically derived models. CoMFA, however, appreciably outperforms EVA with both the oxadiazole dataset, as discussed above, and the nitromethylene heterocycles, although EVA does provide adequate models ($q^2 \geq 0.57$) with the latter dataset. In support of EVA it should be noted that the 11 datasets utilised herein were not chosen at random from the QSAR literature but were selected *because* good CoMFA models were available; i.e. high correlations between the CoMFA descriptors and the relevant biological activity had previously been established. We are thus comparing EVA to CoMFA for a modelling capability for which CoMFA has previously been shown to perform well. It is thus of interest to apply EVA to datasets for which CoMFA fields alone have been found to provide poor correlations with activity. For example, McFarland [43] found it necessary to supplement CoMFA steric and electrostatic fields with a log P descriptor in order to model satisfactorily a set of anticoccidial triazines; studies such as this are the object of future work with EVA. The modelling and prediction of log P values for the structurally diverse BC(DEF) dataset (described above [21]) indicates that EVA can provide both descriptive and predictive models for a set of diverse structures for which mutual alignment would be difficult if not impossible.

The problems associated with alignment in CoMFA have already been noted. However, an advantage of alignment (and the nature of the field-based descriptors themselves) is that it provides a powerful means of visualising the results of an analysis, in the form of graphical displays of the structural regions indicated to be most highly positively or negatively correlated with binding activity. Nonetheless, it should also be pointed out that, despite the undoubted utility of the CoMFA graphical display, it does not indicate to the user precisely which atoms are responsible for the modelled correlations since the grid-based descriptors themselves consist of contributions from more than one atom. These contributions, of course, tend to be dominated by the influence of a single atom when calculating potentials sufficiently close to a given atom. However, at other positions in 3D space interpretation is less clear, particularly with electrostatic fields where there is a 1/r distance-dependence function rather than the $1/r^6$ dependence usually associated with the attractive part of the Lennard-Jones function. As a result, a technique known as PAC (predicted activity contributions) [44] for assessing activity contributions has recently been developed. An analogous problem arises with EVA inasmuch as, depending on the kernel width, very many of the EVA variables for a given structure may consist of contributions from more than one normal mode frequency. Therefore, decomposition, in terms of activity contributions, is required together with subsequent interpretation of the normal mode itself. Work is currently underway to develop such procedures and promises to make EVA a powerful QSAR technique.

Conclusions

It has been demonstrated that EVA can be used to develop good predictive QSAR models for a range of datasets exhibiting biological endpoints ranging from receptor–ligand affinities to those more closely related to transport and distribution effects. This has been achieved without the need to align the structures concerned. These results are all the more encouraging given that the datasets were selected solely on the basis that good CoMFA models were available. At this stage EVA has been evaluated in terms of LOO cross-validation statistics alone, and a subsequent paper [37] addresses the question of more rigorous validation of EVA QSAR through the application of separate test set predictions and extensive random permutation testing. This following paper also considers the importance of conformation selection in relation to an EVA QSAR study.

Of the three means investigated for calculating normal modes, the more intensive semiempirical methods AM1 and PM3 appear to outperform AMBER-based modelling, but not in all cases. As with all such techniques, the method of choice for a given problem should be related to the known chemical or computational merits or deficiencies of the modelling tool.

A wide-ranging evaluation of the main EVA parameters has been performed, resulting in the determination of upper bound values for the sampling increment in relation to the selected Gaussian spread; this ensures that there is no information loss prior to PLS analysis. The effect of changes to the Gaussian standard deviation on the resulting QSAR models has been systematically analysed. In most, but not all, cases, there appears to be no great difference in LOO q^2 scores where models have been developed using a wide range of σ terms. However, where there are significant differences related to the choice of σ it is not possible to recommend the use of any single σ value for QSAR modelling purposes; the use of a low (say, 5 cm^{-1}), medium (say, 10 cm^{-1}), high (say, 25 cm^{-1})

and possibly a very high (say, 50 cm^{-1}) σ value would seem to be sufficient to cover the performance range observed in the studies described in this article. The future development of EVA is to include enhancement of the standardisation procedure technique (localised Gaussian σ) and the visualisation of structural features (contra-)indicated by an EVA model.

Acknowledgements

We thank Shell Research Limited, the Science and Engineering Research Council and the Biotechnology and Biological Research Council for funding, Tripos Inc. for the provision of hardware and software, and the following for providing the QSAR datasets: Shell Research Limited; Y.C. Martin and K.H. Kim (Abbott Laboratories); C.L. Waller (University of North Carolina); C. Silipo and A. Vittoria (Università di Napoli); J. McFarland (Pfizer Inc.); and K.F. Koehler (Searle Research and Development). This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

1 Hansch, C. and Fujita, T., J. Am. Chem. Soc., 86 (1964) 1616.
2 Ghose, A.K. and Crippen, G.M., Mol. Pharmacol., 37 (1990) 725.
3 Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
4 Doweyko, A.M., J. Med. Chem., 31 (1988) 1396.
5 Wiese, M., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 431–442.
6 Kim, K.H. and Martin, Y.C., J. Org. Chem., 56 (1991) 2723.
7 Kim, K.H., Greco, G., Novellino, E., Silipo, C. and Vittoria, A., J. Comput.-Aided Mol. Design, 7 (1993) 263.
8 Kellogg, G.E., Semus, S.F. and Abraham, D.J., J. Comput.-Aided Mol. Design, 5 (1991) 545.
9 Karelson, M., Lobanov, V.S. and Katritzky, A.R., Chem. Rev., 96 (1996) 1027.
10 Rhyu, K.-B., Patel, H.C. and Hopfinger, A.J., J. Chem. Inf. Comput. Sci., 35 (1995) 771.
11 Hansch, C. and Leo, A., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. ACS Professional Reference Book, American Chemical Society, Washington, DC, U.S.A., 1995.
12 Klebe, G., Abraham, U. and Mietzner, T., J. Med. Chem., 37 (1994) 4130.
13 Cho, S.J. and Tropsha, A., J. Med. Chem., 38 (1995) 1060.
14 Kroemer, R.T. and Hecht, P., J. Comput.-Aided Mol. Design, 9 (1995) 205.
15 Kroemer, R.T. and Hecht, P., J. Comput.-Aided Mol. Design, 9 (1995) 396.
16 Silverman, B.D. and Platt, D.E., J. Med. Chem., 39 (1996) 2129.
17 Muresan, S., Sulea, T., Ciubotariu, D., Kurunczi, L. and Simon, Z., Quant. Struct.–Act. Relatsh., 15 (1996) 31.
18 Wagener, M., Sadowski, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
19 Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. and Gramatica, P., J. Chemosphere, 32 (1996) 1527.
20 Jonathan, P., McCarthy, W.V. and Roberts, A.M.I., J. Chemometrics, 10 (1996) 189.
21 Ferguson, A.M., Heritage, T., Jonathon, P., Pack, S.E., Phillips, L., Rogan, J. and Snaith, P.J., J. Comput.-Aided Mol. Design, 11 (1997) 143.
22 Ginn, C.M.R., Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W., J. Chem. Inf. Comput. Sci., 37 (1997) 23.
23 Herzberg, G., Molecular Spectra and Molecular Structure. II. Infrared and Raman Spectra of Polyatomic Molecules, 8th ed., Van Nostrand Company, Inc., New York, NY, U.S.A., 1945.
24 Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 523–550.
25 Wold, S., In Van de Waterbeemd, H. (Ed.) Methods and Principles in Medicinal Chemistry, Vol. 2, Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1995, pp. 195–218.
26 Cramer, R.D., J. Am. Chem. Soc., 102 (1980) 1837.
27 SYBYL, Tripos Associates Inc., St. Louis, MO, U.S.A.
28 Carroll, F.I., Gao, Y.G., Rahman, M.A., Abraham, P., Parham, K., Lewin, A.H., Boja, J.W. and Kuhar, M.J., J. Med. Chem., 34 (1991) 2719.
29 Jain, A.N., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.
30 Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2091.
31 Good, A.C., So, S.-S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.
32 ASP, 1993, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, U.K.
33 Waller, C.L. and McKinney, J.D., J. Med. Chem., 35 (1992) 3660.
34 Stewart, J.J.P., J. Comput.-Aided Mol. Design, 4 (1990) 1.
35 Quantum Chemistry Program Exchange (QCPE), University of Indiana, Bloomington, IN, U.S.A.
36 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S. and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.
37 Turner, D.B., Willett, P.W., Ferguson, A.M. and Heritage, T., J. Med. Chem., submitted.
38 Cruciani, G., Clementi, S. and Baroni, M., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 551–564.
39 Kubinyi, H. and Abraham, U., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 717–728.
40 Topliss, J.G. and Edwards, R.P., J. Med. Chem., 22 (1979) 1238.
41 Turner, D.B., Ph.D. Thesis, Sheffield University, Sheffield, U.K., 1996.
42 Wold, S. and Eriksson, L., In Van de Waterbeemd, H. (Ed.) Methods and Principles in Medicinal Chemistry, Vol. 2, Chemometric Methods in Molecular Design, VCH, Weinheim, Germany, 1993, pp. 309–318.
43 McFarland, J.W., J. Med. Chem., 35 (1992) 2543.
44 Waszkowycz, B., Clark, D.E., Frenkel, D., Li, J., Murray, C.W., Robson, B. and Westhead, D.R., J. Med. Chem., 37 (1994) 3994.
45 Allen, M.S., Laloggia, A.J., Dorn, L.J., Martin, M.J., Costantino, G., Hagen, T.J., Koehler, K.F., Skolnick, P. and Cook, J.M., J. Med. Chem., 35 (1992) 4001.
46 Greco, G., Novellino, E., Silipo, C. and Vittoria, A., Quant. Struct.–Act. Relatsh., 10 (1991) 289.