

## Self-Organizing Molecular Field Analysis: A Tool for Structure–Activity Studies

Daniel D. Robinson, Peter J. Winn, Paul D. Lyne, and W. Graham Richards\*

*Physical and Theoretical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, United Kingdom*

*Received June 29, 1998*

Self-organizing molecular field analysis (SOMFA) is a novel technique for three-dimensional quantitative structure–activity relations (3D-QSAR). It is simple and intuitive in concept and avoids the complex statistical tools and variable selection procedures favored by other methods. Our calculations show the method to be as predictive as the best 3D-QSAR methods available. Importantly, steric and electrostatic maps can be produced to aid the molecular design process by highlighting important molecular features. The simplicity of the technique leaves scope for further development, particularly with regard to handling molecular alignment and conformation selection. Here, the method has been used to predict the corticosteroid-binding globulin binding affinity of the “benchmark” steroids, expanded from the usual 31 compounds to 43 compounds. Test predictions have also been performed on a set of sulfonamide endothelin inhibitors.

### Introduction

Quantitative structure–activity relations (QSAR) and three-dimensional quantitative structure–activity relations (3D-QSAR) have had a profound impact on medicinal chemistry.<sup>1–4</sup> The ability to produce quantitative correlations between three-dimensional properties of molecules and the biological activity of these compounds is of inestimable value in deciding upon the choice of future synthetic chemistry.

Pre-eminent among the techniques used is comparative molecular field analysis (CoMFA) introduced by Cramer.<sup>5</sup> As in the related GRID technique of Goodford,<sup>6</sup> molecules are sited in a three-dimensional grid. At each grid point an interaction energy related to shape or electrostatic potential is calculated. For a series of molecules, the biological activity is related to the set of interaction energies which may run into many thousands of variables. Partial least squares analysis<sup>7</sup> (PLS) is then used to extract the relationship between the interaction energies and the biological activity. The problem is greatly underdetermined, but PLS produces the underlying relationship by reducing the dimensionality of the descriptor space in a way that leaves the most significant contributions to the correlation. Too much insignificant data or noise can degrade the statistical quality of the QSAR.<sup>8</sup> For instance, making the grid too fine may lower the quality of the correlation.<sup>9</sup> To minimize such problems, techniques have been developed to filter data points prior to the PLS step.<sup>10,11</sup> For example, grid points where the calculated interaction energy is low may be omitted. Results are also improved by variable selection procedures<sup>12–17</sup> and by grouping points.<sup>18</sup>

Molecular similarity was introduced as a concept by Carbo.<sup>19</sup> The use of similarity as a 3D-QSAR tool was introduced by Good et al.<sup>9</sup> and used by several other groups.<sup>15,20,21</sup> Similarity between pairs of molecules can be defined in terms of shape or of electrostatic potential, but instead of a large number of values at grid points

we have a single numerical measure of overall similarity. A set of molecules may be compared to a single reference molecule to yield a predictive QSAR.<sup>22–25</sup> More information can be obtained from a matrix of the similarity indices between all pairs of molecules in a training set using PLS to derive a QSAR.<sup>9,26,27</sup> Statistically, similarity matrix correlations are comparable with those from CoMFA<sup>9,15</sup> but, while they have the advantage of much smaller data matrices than those of CoMFA, there is the corresponding loss of a graphical display of significant features of the molecules.



Our new SOMFA technique has similarities to both techniques, but has the advantage of its inherent simplicity. It has affinities with the Free–Wilson method<sup>28</sup> and also the work of Doweyko<sup>29–32</sup> but again is conceptually simpler and more comprehensive. Like CoMFA, a grid-based approach is used. However no probe interaction energies need to be evaluated. Like the similarity methods it is the intrinsic molecular properties, such as shape, that are used to develop the QSAR models. Further, because of its inherent simplicity we believe the method has great potential for development, particularly in regard to the alignment and conformational problems inherent in 3D-QSAR. Ongoing work in this area shows considerable promise.

### Methods

**The SOMFA Methodology.** As with all QSAR techniques a model is built from a set of molecules of known activity; these molecules constitute the training set. Crucial to SOMFA is the notion of the “mean centered activity”. By subtracting the mean activity of the molecular training set from each molecule's activity, we obtain a scale where the most active molecules have positive values and the least active molecules have negative values.

Three-dimensional grids are created as in other QSAR techniques with values at the grid points representing the shape or electrostatic potential. Shape values are given a value of 1 inside the van der Waals envelope, 0 outside. Electrostatic potential values at grid points are calculated in the normal

**Table 1.** Example of a Set of Molecules of Varying Activity, Together with Their Mean Centered Binding Affinities, for a Hypothetical Active Site

structure	binding affinity	mean centered binding affinity
	1	-1
	2	0
	3	1

manner from the partial charges distributed across the atom centers. The most important step is that the value of the shape or electrostatic potential at every grid point for a given molecule is multiplied by the mean centered activity for that molecule. This weights the grid points so that the most active and least active molecules have higher values than the less interesting molecules close to the mean activity. It thus acts as a form of descriptor filtering.

In general, a SOMFA grid can be trained on any calculable molecular property. The grids for each molecule in the training set are combined to give master grids for each property. The value of a SOMFA master grid point at a given  $x,y,z$  is defined by

$$\text{SOMFA}_{x,y,z} = \sum_i^{\text{Training\_Set}} \text{Property}_i(x,y,z) \text{Mean\_Centred\_Activity}_i \quad (1)$$

In the examples presented here, the mean centered activity was represented on a logarithmic scale. The values at each point of a property master grid can be displayed to highlight features favorable or unfavorable to activity. For example, the shape master grid is a template of the areas of steric bulk which enhance or detract from activity.

A QSAR relating a property, such as shape, to molecular activity can be derived from the property master grid in the following way. For every molecule ( $i$ ) in the training and prediction set, an estimate of the activity of the molecule as defined by a certain property can be obtained by using eq 2.

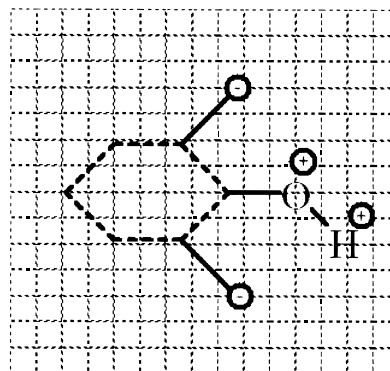
$$\text{SOMFA}_{\text{property},i} = \sum_x \sum_y \sum_z \text{Property}_i(x,y,z) \text{SOMFA}_{x,y,z} \quad (2)$$

Linear regressions between the  $\text{SOMFA}_{\text{property},i}$  values and the logarithms of the experimental activities for the training set are then derived. Calculating the correlation coefficient indicates the potential importance of a given property. The linear equations produced can be used to predict the activity of compounds in the prediction set from their  $\text{SOMFA}_{\text{property},i}$  values. A better method is to combine the predictive power of the different  $\text{SOMFA}_{\text{property},i}$ . Here we combined the individual property predictions using a weighted average of the shape and electrostatic potential based QSAR, using a mixing coefficient ( $c_1$ ) as in eq 3.

$$\text{Activity} = c_1 \text{Activity}_{\text{Shape}} + (1 - c_1) \text{Activity}_{\text{ESP}} \quad (3)$$

Clearly multiproperty predictions could have been obtained through multiple linear regression. Using eq 3 instead gives greater insight into the resultant model by allowing the study of the variation in predictive power with different values of  $c_1$ .

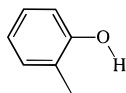
**A Simplified Example.** Table 1 shows the structures of three molecules together with their binding affinities for a

**Figure 1.** The distribution of the values from the training of the SOMFA lattice for the compounds in Table 1. The broken lines show the outline of the original molecules.

hypothetical binding site. The example is set up such that a hydroxyl group enhances binding and excess steric bulk adjacent to the hydroxyl group, associated with the methyl groups in compounds **1** and **2**, reduces binding. The first step in the SOMFA process is to calculate the mean centered affinities/activities for our known molecules. In doing this we achieve a scale of activity where all high activity compounds have a positive activity, while all low activity compounds have a negative activity. The molecules are aligned by superimposing the rings, as in Table 1. We now create a grid around them. In our simplified example we add the mean centered binding affinity (defined in Table 1) of each molecule to every grid point crossed by the molecule. As shown in Figure 1, this gives a trained grid based on molecular shape. The common OH group present in the two more active molecules is left with a net positive grid value associated with it. The common benzenoid core of all the molecules has no net value associated with it, and the methyl groups have a negative grid value associated with them. Thus the mean centered activities have been used as a filtering mechanism to highlight the features which differentiate the high-affinity and low-affinity compounds. An important point is that the benzene ring is not found to be significant. This set of three is not diverse enough for any conclusion to be made about this structural feature, and SOMFA correctly highlights this and filters it from the analysis. The final result is a grid-based map that can be used to aid molecular design of compounds with enhanced binding affinity for our hypothetical site. Since the method is grid-based, there is no necessity for the compounds under study to be structural analogues, only that they can be suitably aligned. We would expect any number of diverse compounds with a common binding site to be suitable for treatment by this method.

In general we expect that high-activity compounds with common structural features (i.e., a pharmacophore) would overlay these features at the same point on a master grid. The grid values from successive high-activity compounds would reinforce each other leading to a final master grid with positive value associated with the features common to these high-activity compounds. Similarly low-activity compounds would also be expected to have some common structural features that lead to a build up of negative grid values. Since the grid values are assigned according to mean centered activity, compounds with intermediate activity will have little effect on the final grid values. It is clear from this description that overly small data sets will not produce the overlapping of features required for SOMFA. The quality of the model does improve rapidly with data size, and this is not expected to be a problem for normal size QSAR data sets (10 or more compounds).

Returning to our hypothetical example, the trained grid can be used to predict the activity of a novel compound such as the one in Figure 2. When the structure is overlaid on the trained SOMFA lattice, we can see that the molecule is associated with only one area of poor steric overlap while simultaneously possessing the hydroxyl group which we have



**Figure 2.** An example of an unknown molecule to be predicted by SOMFA.

identified as being the most important feature for activity. From a summation of all of these good and poor overlaps, SOMFA would conclude that methyl-phenol would have an affinity for the binding site between 2,6-dimethyl phenol and phenol. This value is intuitively in agreement with our knowledge of the hypothetical binding site.

**The SOMFA Code.** The initial version of SOMFA has been coded to have the following features, which have been utilized to produce the results in this article. Molecular alignment can be performed using a principal component analysis method (PCA)<sup>33,34</sup> to align common molecular cores defined by the user. Code was also developed to automatically set up SOMFA grids for each molecule in the prediction set for either, or both, electrostatics and shape. Combining individual molecular grids can produce master grids, and these can be visualized via a graphical interface. The interface was designed to display only those grid points with values above a threshold level defined by a hard-coded function, thus displaying only the maxima and minima of the master grids. The SOMFA<sub>Property</sub> values, SOMFA<sub>ESP</sub> and SOMFA<sub>shape</sub>, were evaluated by the SOMFA code and written to disk. All linear regression and correlation calculations were performed in Excel 5.0.

The typical resources required by the code can be determined from the following example. SOMFA grids with 2 lattice points per angstrom over a 25 Å cube used approximately 1.5 MB of storage per molecule. This is well within the capabilities of a modern workstation PC. Calculations on such a lattice took approximately 30 s on a dual processor 300 MHz Pentium II under Microsoft Windows NT4.0.

The PCA alignment technique used here is only suitable for structural analogues, with a readily identifiable structural motif. The common structural motif is defined by the user, and the coordinates of each atom in the motif are stored in three vectors,  $\mathbf{X}_i$ ,  $\mathbf{Y}_i$ , and  $\mathbf{Z}_i$ , for each molecule,  $i$ . The centroid of the motif for each molecule is then shifted to the origin by subtracting the mean  $\mathbf{X}_i$ ,  $\mathbf{Y}_i$ ,  $\mathbf{Z}_i$  from all the atom coordinates in that molecule. To align the resulting positions  $\mathbf{X}'_i$ ,  $\mathbf{Y}'_i$ ,  $\mathbf{Z}'_i$  of the common motif in all the molecules, a rotation is now required. This is derived by finding the eigenvectors of the autocovariance matrix of the vectors  $\mathbf{X}'_i$ ,  $\mathbf{Y}'_i$ ,  $\mathbf{Z}'_i$ . These eigenvectors are the principal components of the molecule  $i$ . The autocovariance matrix  $\mathbf{S}_i$  for a molecule is defined as

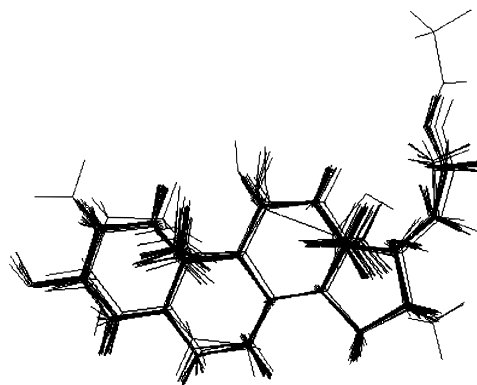
$$\mathbf{S}_i = \begin{bmatrix} \mathbf{X}'_i \cdot \mathbf{X}'_i & \mathbf{X}'_i \cdot \mathbf{Y}'_i & \mathbf{X}'_i \cdot \mathbf{Z}'_i \\ \mathbf{X}'_i \cdot \mathbf{Y}'_i & \mathbf{Y}'_i \cdot \mathbf{Y}'_i & \mathbf{Y}'_i \cdot \mathbf{Z}'_i \\ \mathbf{X}'_i \cdot \mathbf{Z}'_i & \mathbf{Y}'_i \cdot \mathbf{Z}'_i & \mathbf{Z}'_i \cdot \mathbf{Z}'_i \end{bmatrix} \quad (4)$$

The eigenvectors of  $\mathbf{S}_i$  are denoted as  $\lambda_{i,0}$ ,  $\lambda_{i,1}$ , and  $\lambda_{i,2}$  in descending order of eigenvalue. We can rotate the molecule by forming the rotation matrix  $\Lambda_i$  for each molecule,  $i$ , and applying this matrix to all of the atomic positions in the molecule.

$$\Lambda_i = \begin{bmatrix} \lambda_{i,0} \\ \lambda_{i,1} \\ \lambda_{i,2} \end{bmatrix} \quad (5)$$

Each molecule,  $i$ , is rotated by its associated  $\Lambda_i$ , yielding coincident atomic positions in the common motif. It is important to check that the determinant of  $\Lambda_i$  is equal to unity otherwise  $\Lambda_i$  may invert the molecule, since if  $\lambda_{i,0}$  is an eigenvector of  $\Lambda_i$  so is  $-\lambda_{i,0}$ .

**Testing SOMFA.** Two sets of molecules were investigated to test the SOMFA technique. The first set was the widely studied steroid set<sup>5,9,17,35-47</sup> which has become a QSAR bench-



**Figure 3.** Overlay of all 31 steroid structures using PCA.

mark. These steroids exhibit a range of binding affinities for corticosteroid-binding globulin (CBG). The second set was a series of sulfonamide compounds<sup>48</sup> which act as endothelin inhibitors. Both sets of molecules were geometry-optimized at the AM1<sup>49</sup> level using MOPAC 6.0.<sup>50</sup> The atom-centered point charges used to generate the electrostatic potential in SOMFA were derived by fitting to reproduce the AM1 quantum mechanical molecular electrostatic potential, using the RATTLE software.<sup>51,52</sup>

The steroid compounds were aligned by forming a PCA fit of the common steroid core across all molecules. The alignment of the molecules is shown in Figure 3. We should bear in mind that this alignment does not take into account any electrostatic features; consequently, we may be biasing the results in favor of shape. A grid size of 25 Å<sup>3</sup> was used for all the steroid molecules. This allowed the value of the electrostatic potential to fall to a low level before grid truncation. For the results presented here, a grid resolution of 2 points per angstrom was used.

In line with most published results for the steroid set (Table 2), the relative binding affinities for CBG of molecules **1–21** were used to train the SOMFA grid. Molecules **22–31** formed prediction set A and were then predicted from this grid; molecule **31** is a significant outlier for this set. It has been suggested that this anomaly is due to it being fluorinated at the 9 carbon<sup>5,39,43</sup> (see Table 2a, steroid **1** for the standard numbering system of the steroid ring); however, we would argue that a QSAR method should be robust enough to handle such a small change of structure. Referring back to the original experimental work,<sup>53-55</sup> a better reason for this anomaly would be the different experimental techniques used to calculate the CBG binding affinities. Indeed this is even alluded to by Westphal.<sup>53</sup> For this reason we have also used an additional 12 steroids,<sup>56</sup> where CBG binding affinities have been obtained by the same experimenters, using the same technique as that used for the 21 training steroids. These extra molecules include a compound fluorinated at the same position as the notorious steroid **31**. These steroids will be referred to as steroid prediction set B.

The 35 sulfonamide molecules (Table 3) and associated data were taken from a paper by Krystek et al.<sup>48</sup> The structures were aligned using PCA; the common structural motif used is denoted by the asterisks on structure **1** of Table 3. The alignments are shown in Figure 4. For the SOMFA analysis a 30 Å cubic grid with 2 points per angstrom was used. The training set was made up from every even numbered sulfonamide (presented in Table 3). The other half of the molecular set was used for test predictions. This represents a very harsh test of the predictive ability of SOMFA.

## Results and Discussion

The results of grid-based QSAR are usually dependent on the grid resolution and the orientation of the molecules with respect to the grid. For example, it is known that in CoMFA if too coarse a grid is used, the

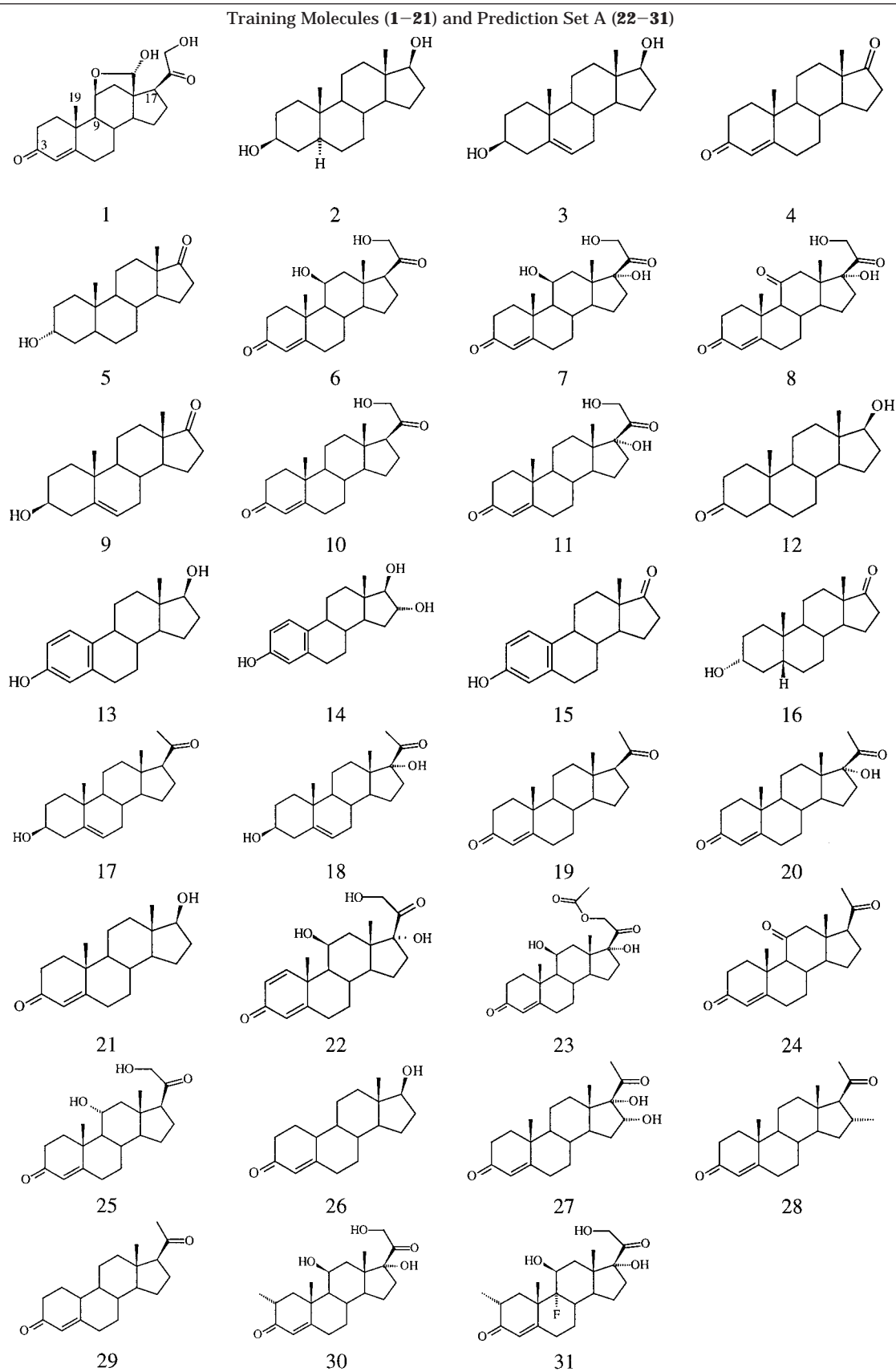
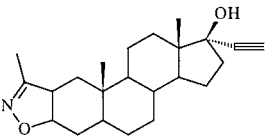
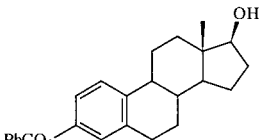
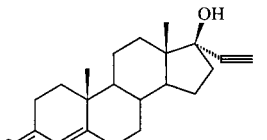
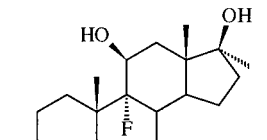
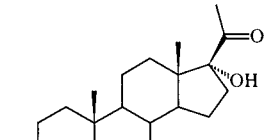
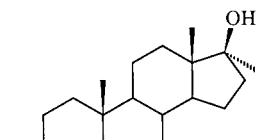
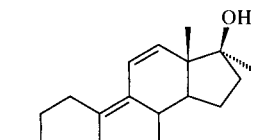
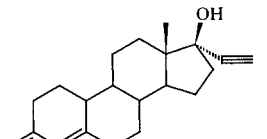
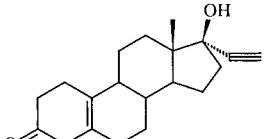
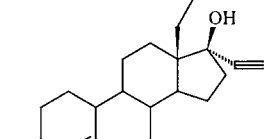
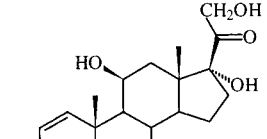
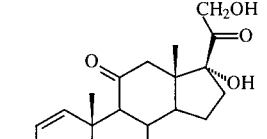
**Table 2.** Structures of the Steroid Compounds



Table 2 (Continued)

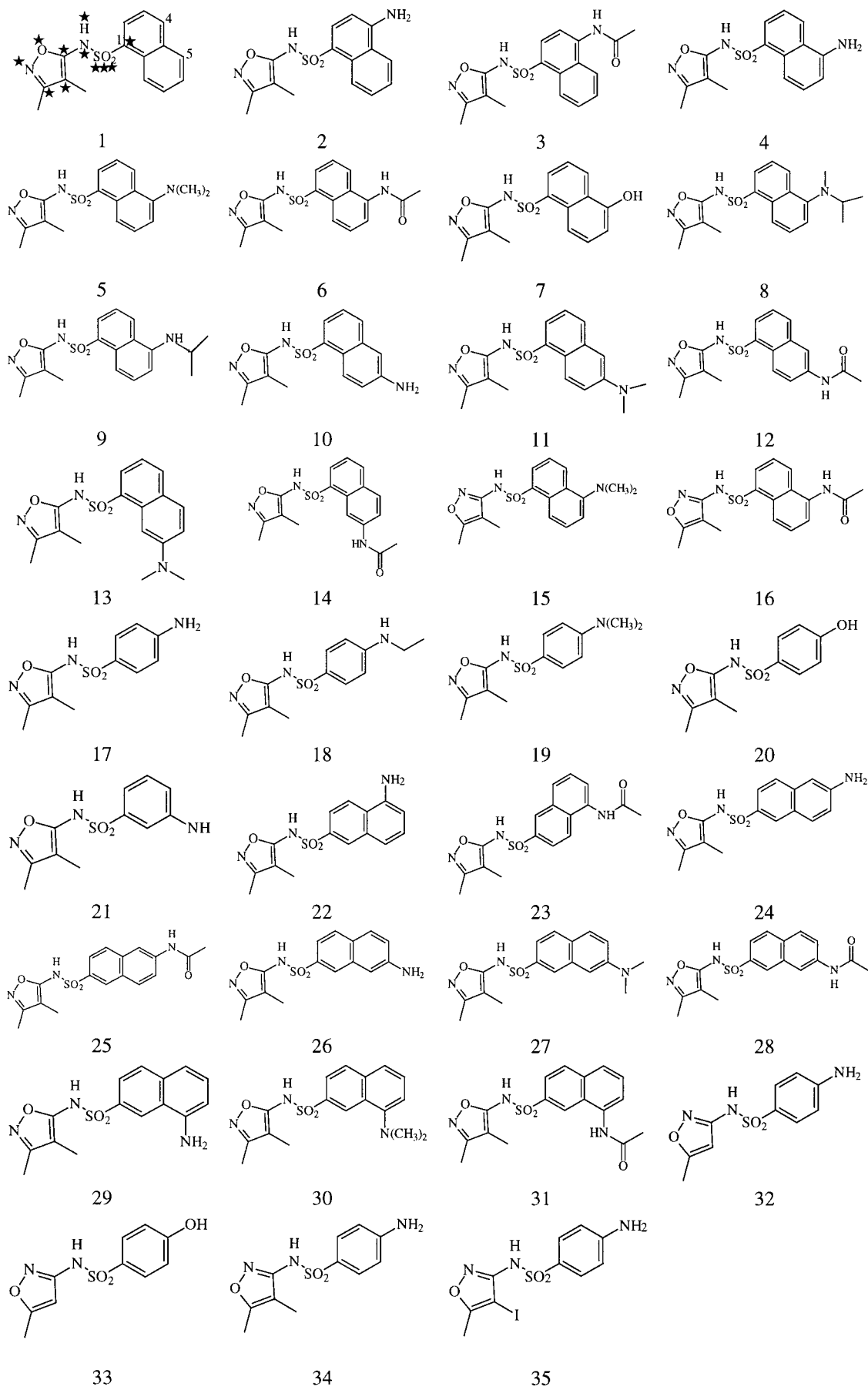
Prediction Set B			
			
32 Danazol	33 Estradiol benzoate	34 Ethisterone	35 Fluoxymesterone
$\log(k) = -6.813$	$\log(k) = -5.0$	$\log(k) = 5.322$	$\log(k) = 5.0$
			
36 Medroxyprogesterone	37 Methyltestosterone	38 Methyltrienolone	39 Norethindrone
$\log(k) = 6.908$	$\log(k) = 5.0$	$\log(k) = 5.362$	$\log(k) = 5.255$
			
40 Norethynodrel	41 Norgestrel	42 Prednisolone	43 Predinsone
$\log(k) = 5.0$	$\log(k) = 5.0$	$\log(k) = 7.613$	$\log(k) = 6.505$

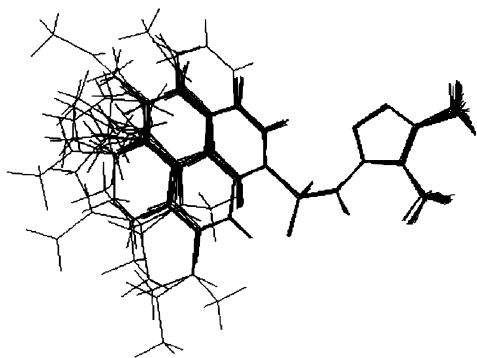
quality of QSAR is likely to be degraded; however, if the grid is too fine, the amount of noise in the descriptor data is increased and thus the model is again degraded.<sup>8,9</sup> During the SOMFA investigation of the steroid test set, grid spacings of 1, 0.5, 0.25, and 0.2 Å were investigated. The 1 Å grid spacing produced a good correlation. This improved marginally with the 0.5 Å spacing used for the results presented here. Further increases in resolution produced further small increases in model quality but not enough to warrant the extra computational time. The relative orientation of the molecules with respect to the grid were not explicitly investigated, since marginal variation of results with grid spacing suggest this will not be significant at the 0.5 Å resolution used here. The results using this grid resolution are discussed below.

The value of the correlation coefficients, presented in Figures 5–8, for training sets and prediction sets are excellent in both the steroid and sulfonamide cases. The ability to predict on novel data is the most important test for any new QSAR method. By using test sets to make genuine but verifiable predictions we have shown the high-correlation coefficients of the training sets are a true reflection of the models' predictive ability and not due to chance correlations.

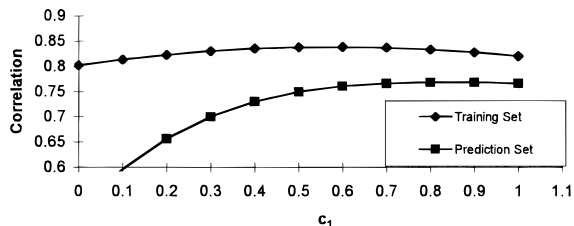
The correlation coefficient between predicted activity and experimental activity varies with  $c_1$  (see eq 3), as shown in Figures 5 and 6 for the steroid and sulfonamide sets, respectively. We note the maximum of correlation for the training set curves in both figures does not correspond exactly to the maximum of correlation for the prediction set curves. However, the maxima for the training sets are close to the maxima in the prediction set curves. Indeed, predictions made at the optimal  $c_1$  defined by the training data have almost identical correlation coefficients to predictions made at the maxima of the prediction sets. It thus seems reasonable to presume that the value of  $c_1$  that yields optimal predictive power in the training set is a reasonable value to use for true predictions. The optimal training ratio of 6:4 in favor of shape was used for all steroid prediction set results discussed later. Similarly, the optimal training ratio of approximately 7:3 in favor of shape was used for all sulfonamide prediction set results. This ensures that subsequent comparisons with other QSAR methods are fair.

Looking at Figure 7, the steroid set predicted activities versus experimental activities, steroid **31** was predicted as an outlier, as with most other methods. However in set B we see steroid **35**, the similarly fluorinated

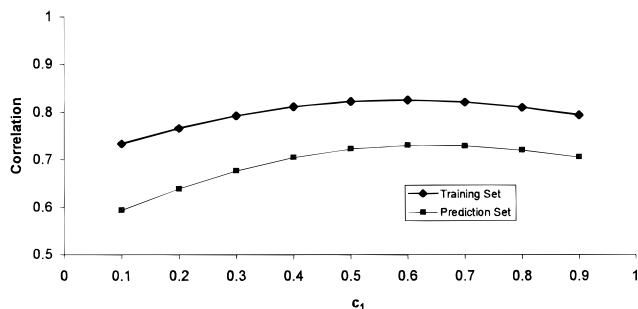
**Table 3.** Structure of the Sulfonamides



**Figure 4.** Overlay of all 35 sulfonamide structures using PCA.



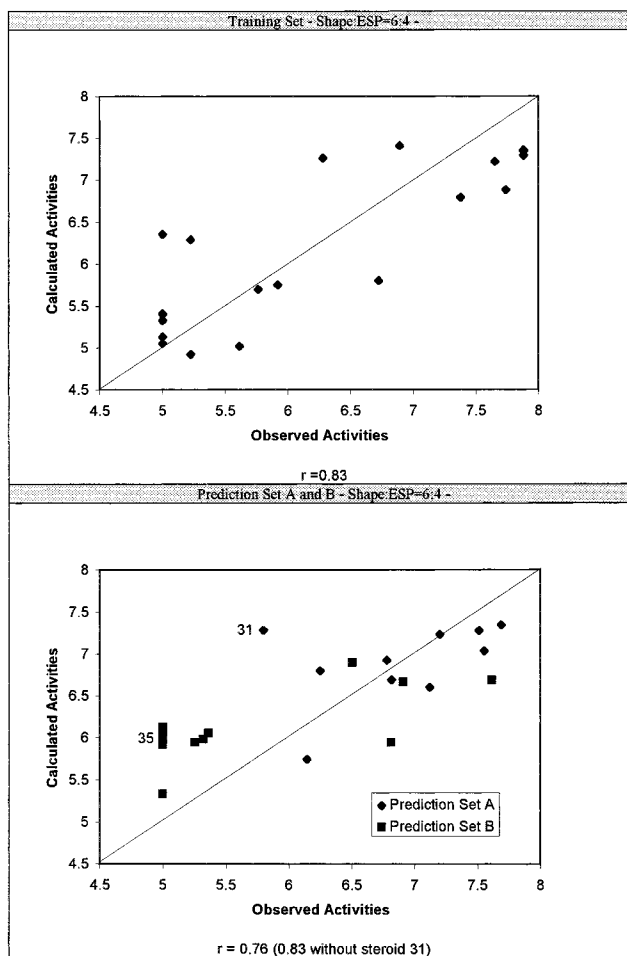
**Figure 5.** The variation with  $c_1$  of the correlation between predicted and experimental binding affinities for the steroid data set. The prediction set data includes all set A and all set B.



**Figure 6.** The variation with  $c_1$  of the correlation between predicted and experimental endothelin inhibition for the sulfonamide data set.

compound, is not such an outlier. This suggests that fluorination per se is not a problem for the steroid set. The problems with steroid 31 are probably due to the differences in the experimental techniques used to derive the training data and those used to derive the data for steroids **22–31**.

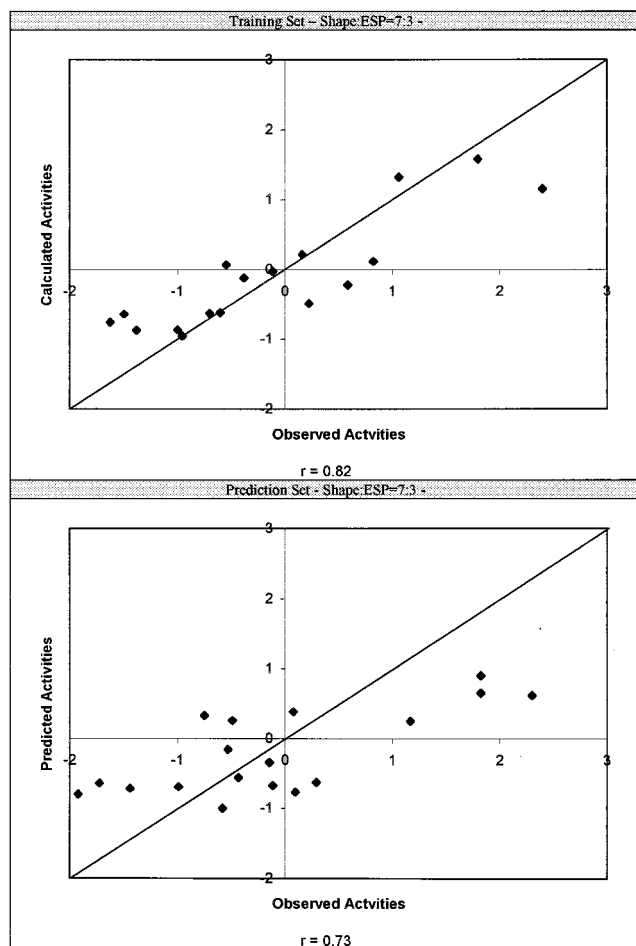
A comparison of the predictive powers of SOMFA compared with other QSAR methods is given in Table 4. The table shows two sets of CoMFA results, derived by Bravi et al.<sup>41</sup> These results improve on the original CoMFA paper<sup>5</sup> through the use of conformational selection techniques and a similarity-based alignment algorithm.<sup>39</sup> The better CoMFA results, CoMFA (FFD) were obtained using a fractional factorial design (FFD),<sup>41</sup> using only those training data found to have the most explanatory power. The other results used for comparison were obtained using the neural network based COMPASS technique,<sup>38</sup> the similarity matrix technique,<sup>57</sup> and the MS-WHIM<sup>41</sup> approach of looking at the molecular surface parameters of a molecule using an alignment invariant method. The quality of the models is to be compared using the so-called standard deviation



**Figure 7.** The scatter of the observed against predicted binding affinities for the steroid data using a 6:4 weighting of shape to electrostatic potential. The plotted lines are  $y = x$ .

of errors of prediction (SDEP),<sup>41</sup> which is the root-mean-square error of the predictions. For the full prediction set A, of 10 molecules, SOMFA yields the best predictions, as judged by the SDEP. When steroid **31** is ignored, COMPASS has the lowest SDEP; however, SOMFA and CoMFA (FFD) have comparable predictivities to COMPASS and have the advantage that the results can be more easily interpreted. We also note that the SOMFA method has not used any form of fractional factorial design and is considerably simpler.

The visualization of the shape master grid for the steroid sets is shown in Figure 9. Steroid 1 is included for reference. In this map of important features we see a high density of high-value red points around the side chain at carbon atom 17. We also see a high density of high-value red grid points around the carbon atom 19 methyl group. This suggests both these moieties are sterically favorable, and further steric bulk here may increase CBG binding affinities. We note that these two areas are very similar to the areas favoring steric bulk which were predicted in the original CoMFA paper. Similarly areas where steric bulk is unfavorable for steroid CBG affinity (areas colored blue in Figure 9) are also in similar regions to those highlighted by CoMFA. However, we note that in the SOMFA model these areas have a low density of points, and may not be as important as other molecular features highlighted by the model.



**Figure 8.** The scatter of the observed against predicted activities for the sulfonamide data using a 7:3 weighting of shape to electrostatic potential. The plotted lines are  $y = x$ .

Several views of the electrostatic master grid are given in Figure 10. A full appreciation of the grid can only be obtained from the VDU. SOMFA has no problems identifying the key electrostatic features, which are seen to be consistent with the structural features of the steroids. If one looks at the distribution of hydroxyl and carboxyl groups in the most active (steroids **6** and **7**) and least active (steroids **2**, **3**, **9**, **13**, **14**, **15**) compounds, we see areas where the electrostatic potential might be expected to be important. The SOMFA model finds a large area of negative potential to be important around carbon atom 3. The most active compounds have a carbonyl group at this position. The



**Figure 9.** The steroid set shape master grid. Red represents areas of favorable steric interactions. Blue represents areas of unfavorable steric interactions. Steroid 1 is included as a frame of reference.

model also suggests there is a large, important area of positive potential around the five-membered ring at the other end of the steroid fused ring system. Finally we have a smaller area of negative potential in the vicinity of carbon atom 17. This appears to be associated with a carbonyl group on the side chain of the more active compounds. The carbon 17 side chain on the most active compounds appears as a feature in both the electrostatic and shape grids. Its presence on one grid may be an artifact of its necessity on another grid. Careful choice of ligand design could test this theory, for example by changing the side chain on steroid **6** so that it is only hydrocarbon.

The maps presented fit well with the data provided. They give a good description of the structural features of a steroid that may lead to or detract from activity. This suggests ways of modifying existing steroids to improve their CBG binding affinities. We note that these are maps of the ligand, not the active site. From a map of the ligand we may make inferences about the active site, though this has not been done here. However, it is important to remember that some features that make a ligand active do not necessarily correspond directly to interactions in the binding cleft. There may be features that assist ligand binding in some other way, such as helping the ligand into the binding cleft in the first place.

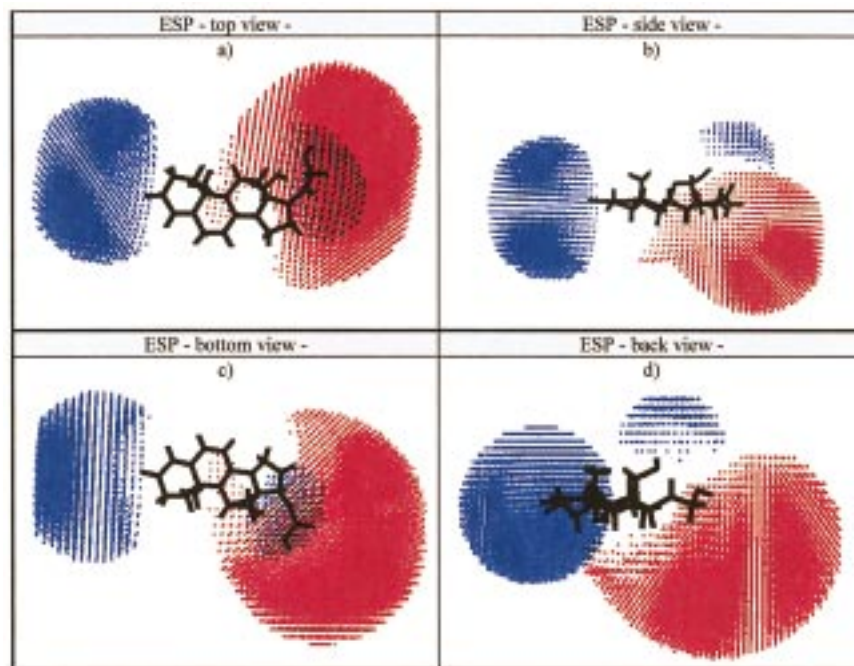
With the sulfonamide set (Figure 8) we see that the calculated activities do not have the same amount of variance as the measured activities. This is a problem of alignment. The aim of this article is to compare SOMFA with established QSAR techniques. Thus to be able to compare the SOMFA master grids with the CoMFA contour plots generated by Krystek et al.<sup>48</sup> we have necessarily used the same alignments. Using a

**Table 4.** Comparison of SOMFA Results with Other Highly Predictive QSAR Techniques for Steroid Prediction Set A

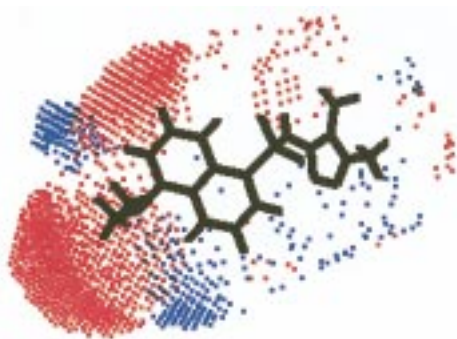
steroid	measured activity	CoMFA <sup>a</sup>	CoMFA (FFD) <sup>a</sup>	similarity matrix analysis <sup>c</sup>	COMPASS <sup>b</sup>	MS-WHIM <sup>a</sup>	SOMFA
<b>22</b>	-7.512	-8.084	-7.883	-7.453	-7.062	-7.300	-7.279
<b>23</b>	-7.553	-7.666	-7.430	-7.022	-7.729	-8.332	-7.034
<b>24</b>	-6.779	-6.538	-6.642	-6.939	-6.462	-6.821	-6.925
<b>25</b>	-7.2	-7.804	-7.705	-7.146	-7.466	-7.445	-7.232
<b>26</b>	-6.144	-6.396	-6.495	-5.908	-5.994	-6.121	-5.744
<b>27</b>	-6.247	-7.346	-6.962	-7.046	-6.383	-6.901	-6.800
<b>28</b>	-7.12	-7.010	-6.848	-6.569	-6.625	-6.532	-6.603
<b>29</b>	-6.817	-6.864	-6.816	-6.850	-7.403	-6.838	-6.692
<b>30</b>	-7.688	-7.970	-7.767	-7.539	-7.741	-7.860	-7.345
<b>31</b>	-5.797	-8.005	-7.793	-7.457	-7.779	-7.491	-7.283
	SDEP	0.837	0.716	0.640	0.705	0.662	0.584
		(0.486) <sup>d</sup>	(0.356) <sup>d</sup>	(0.385) <sup>d</sup>	(0.339) <sup>d</sup>	(0.411) <sup>d</sup>	(0.367) <sup>d</sup>

<sup>a</sup> See ref 39, Table 8. <sup>b</sup> Reference 34, Table 4. <sup>c</sup> Reference 51. <sup>d</sup> Excludes steroid **31**.





**Figure 10.** The steroid set electrostatic potential master grid. Red represents areas where positive potential is favorable, or negative charge is unfavorable. Blue represents areas where negative potential is favorable, or positive charge is unfavorable. Steroid 1 is included as a frame of reference.



**Figure 11.** The sulfonamide set shape master grid. Sulfonamide 5 is included as a frame of reference. A channel of favorable steric interactions runs between two areas of unfavorable steric interactions.



**Figure 12.** The sulfonamide set electrostatic master grid. Sulfonamide 5 is included as a frame of reference. The area of favorable positive potential build lies between the 4 and 5 positions of the 1-substituted naphthyl compounds. The area of negative potential lies above the naphthyl system and the 4 and 5 substituents.

different alignment rule we have obtained significantly improved results with significantly different master grids. To keep this article focused on the SOMFA methodology and how it compares to existing techniques these alternative results will be presented elsewhere. Here we wish to show that for the same alignment as previously used we can get results similar to those from methods already well established. This is another example of how 3D-QSAR methodologies require good structural alignment to obtain the best results.<sup>5,40-42</sup> This is an area where we hope to improve the SOMFA application soon. Indeed we think SOMFA may be ideally suited to tackling the alignment and conformational problems highlighted by Cramer<sup>5</sup> and others<sup>40,41</sup> as the most important stage in 3D-QSAR methods.

Visualization of the sulfonamide trained lattices in terms of shape (Figure 11) and electrostatics (Figure 12) does not show the same level of interpretable detail as the steroid lattices. However, we may identify three regions of interest. First, we can see that steric bulk is tolerated only in a fairly narrow channel from the 5

position of the 1-substituted naphthyl structures. Second, in terms of the electrostatics we can see that a build up of positive charge is apparently favorable below the 4 or 5 positions of the 1-substituted naphthyl structures. Third, we can see a region where addition of negative charge may be expected to enhance activity. All of these features captured by SOMFA were noted by Krystek<sup>48</sup> in his original CoMFA analysis of the data set.

A SOMFA model could be based on any molecular property; here we have used the molecular shape and the molecular electrostatic potential. The SOMFA model also suggests a method of tackling the all-important alignment problem, which all 3D-QSAR methods face. The inherent simplicity of the method allows the possibility of aligning the training compounds as an integral part of the model derivation process and of aligning prediction compounds to optimize their predicted activity. Ongoing work in this area suggests that this new 3D-QSAR technique may yield even better results than those presented here.

## Conclusions

The SOMFA method presented here is a conceptually simple, yet strikingly powerful new QSAR technique. Its predictive power is very good as compared with some of the best methods currently in use, yet it eschews heavy statistical elements. Further, visual maps of the sterically and electrostatically important features of lead compounds can be used to guide the drug design process. The method should be suitable for diverse sets of compounds. Very good results have been obtained for steroids with affinity for CBG and also for a series of sulfonamide endothelin inhibitors. Such is the speed and simplicity of the approach that we believe we can introduce molecular alignment and conformational flexibility into the search for the best 3D-QSAR model.

**Acknowledgment.** D.D.R. is supported by an EPSRC CASE studentship held in conjunction with Oxford Molecular Group PLC, P.W. is funded by an Oxford Molecular Group PLC postdoctoral fellowship. This work was in part supported by the Wellcome Trust. The authors are happy to provide copies of the software. See group Web page <http://bellatrix.pcl.ox.ac.uk>.

## References

- Hansch, C.; Fujita, T. r-s-p analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
- van de Waterbeemd, H. *Advanced computer assisted techniques in drug discovery*; VCH: Weinheim, 1995; Vol. 3.
- van de Waterbeemd, H. *Chemometric methods in molecular design*; VCH: Weinheim, 1995; Vol. 2.
- Kubinyi, H. *3D QSAR in drug design: theory, methods and applications*; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity*; Kulwer Academic Publishers: Dordrecht, The Netherlands, 1998; *3D QSAR in Drug Design: Recent Advances*; Kulwer Academic Publishers: Dordrecht, The Netherlands, 1998.
- Cramer, R. D., III.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
- Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important molecules. *J. Med. Chem.* **1985**, *28*, 849–857.
- Dunn, W. J.; Wold, S.; Edlund, U.; Hellberg, S. Multivariate structure–activity relationships from data from a battery of biological tests and an ensemble of structure descriptors. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131–137.
- Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, P. Predictive ability of regression models part II: selection of the best predictive PLS model. *J. Chemometr.* **1992**, *6*, 347–356.
- Good, A. C.; So, S.-S.; Richards, W. G. Structure–activity relationships from molecular similarity matrixes. *J. Med. Chem.* **1993**, *36*, 433–438.
- Cruciani, G.; Watson, K. A.; Comparative molecular field analysis using GRID force-field and GOLPE variable selection methods in a study of inhibitors of glycogen phosphorylase b. *J. Med. Chem.* **1994**, *37*, 2589–2601.
- Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of comparative molecular field analysis models: effects of data scaling and variable selection using a set of human synovial fluid phospholipase A(2) inhibitors. *J. Med. Chem.* **1997**, *40* (6), 1136–1148.
- Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S.; Generating optimal linear PLS estimations (GOLPE). An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9–20.
- Allen, M. S.; La Loggia, A. J.; Dorn, L. J.; Martin, M. J.; Costantino, G.; Hagen, T. J.; Koehler, K. K.; Skolnick, P.; Cook, J. M. Predictive binding of  $\beta$ -Carboline Inverse Agonists and Antagonists via the CoMFA/GOLPE Approach. *J. Med. Chem.* **1992**, *35*, 4001–4010.
- Cruciani, G.; Clementi, S.; Baroni, M. Variable Selection in PLS Analysis. In *3D QSAR in Drug Design*; Kubinyi, H., Eds; ESCOM: Leiden, 1993; pp 551–564.
- Good, A. C.; Peterson, S. J.; Richards, W. G. QSARs from similarity matrixes. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem.* **1993**, *36*, 2929–2937.
- Cocchi, M.; Menziani, M. C.; De Benedetti, P. G. Use of advanced chemometric tools and comparison of different 3D descriptors in QSAR analysis of prazosin analogues alpha-1-adrenergic antagonists. In *Trends in QSAR and Molecular Modelling 92*; Wermuth, C. G., Ed.; ESCOM: Leiden, 1993; pp 527–529.
- Tominaga, Y.; Fujiwara, I. Prediction-weighted partial least-squares regression method (PWPLS) 2: Application to CoMFA. *J. Chem. Info. Comput. Sci.* **1997**, *37* (6), 1152–1157.
- Pastor, M.; Cruciani, G.; Clementi, S.; Smart region definition: A new way to improve the predictive ability and interpretability of three-dimensional quantitative structure activity relationships. *J. Med. Chem.* **1997**, *40* (10), 1455–1464.
- Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
- Richards, W. G. In *Modeling of biomolecular structures and mechanisms*; Pullman, A., Jortner, A., Pullman, B., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1995; pp 365–369.
- Dean, P. M. In *3D QSAR in drug design: theory, methods and applications*; Kubinyi, H., Ed.; ESCOM Science Publishers B.V.: Leiden, The Netherlands, 1993; pp 150–172.
- Hodgkin, E. E.; Richards, W. G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem., Quantum Biol. Symp.* **1987**, *14*, 105–110.
- Burt, G.; Huxley, P.; Richards, W. G. The Application of Molecular Similarity Calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.
- Seri-Levi, A.; Salter, R.; West, S.; Richards, W. G. Shape Similarity as a Single Independent Variable in QSAR. *Eur. J. Med. Chem.* **1994**, *29*, 687–694.
- Montanari, C. A.; Tute, M. S.; Beezer, A. E.; Mitchell, J. C. Determination of receptor bound drug conformations by QSAR using flexible fitting to derive a molecular similarity index. *J. Comput. Aid. Mol. Des.* **1996**, *10*, 67–73.
- Benigni, R.; Cotta Ramusino, M.; Giorgi, F.; Gallo, G. Molecular similarity matrixes and quantitative structure activity relationships: a case study with methodological implications. *J. Med. Chem.* **1995**, *38*, 629–635.
- Horwell, D. C.; Howson, W.; Higginbottom, M.; Naylor, D.; Ratcliffe, G. S.; Williams, S. Quantitative structure activity relationships (QSARs) of N-terminus fragments of NK1 tachykinin antagonists: a comparison of classical QSARs and three-dimensional QSARs from similarity relationships. *J. Med. Chem.* **1995**, *38*, 4454–4462.
- Free, S. M., Jr.; Wilson, J. W. A Mathematical Contribution to Structure–Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
- Doweyko, A. M. The Hypothetical Active Site Lattice. An Approach to Modelling Active Sites from Data on Inhibitor Molecules. *J. Med. Chem.* **1988**, *31*, 1396–1406.
- Doweyko, A. M.; Mattes, W. B. An application of 3D-QSAR to the analysis of the sequence specificity of DNA alkylation by uracil mustard. *Biochemistry* **1992**, *31* (39), 9388–9392.
- Doweyko, A. M. Three-Dimensional Pharmacophores from Binding Data. *J. Med. Chem.* **1994**, *37*, 1769–1778.
- Kaminski, J. J.; Doweyko, A. M. Anti-ulcer agents. 6. Analysis of the in vitro biochemical and in vivo gastric antisecretory activity of substituted imidazol [1,2-*a*]pyridines and related analogues using comparative molecular field analysis and hypothetical active site lattice methodologies. *J. Med. Chem.* **1997**, *40* (4), 427–436.
- Manly, B. F. J. *Multivariate Statistical Methods. A primer*. Chapman and Hall: London, 1986.
- Chatfield, C.; Collins, A. J. *Introduction to Multivariate Analysis*. Chapman and Hall: London, 1980.
- Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural-networks: Application to the analysis of corticosteroid-binding globulin activity of steroids. *J. Comput. Aid. Mol. Des.* **1996**, *10* (6), 521–534.
- Jain, N. A.; Koile, K.; Chapman, D. COMPASS – Predicting Biological-Activities from Molecular-Surface Properties – Performance Comparisons on a steroid Benchmark. *J. Med. Chem.* **1994**, *37*, 2315–2327.
- Tominaga, Y.; Fujiwara, I. Novel 3D descriptors using excluded volume: Application to 3D quantitative structure–activity relationships. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (6), 1158–1161.
- Lobato, M.; Amat, L.; Besalu, E.; CarboDorca, R. Structure–activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, *16* (6), 465–472.

- (39) Norinder, U. 3D-QSAR investigation of the Tripos benchmark steroids and some protein tyrosine-kinase inhibitors of styrene type using the TDQ approach. *J. Chemometr.* **1996**, *10*, 533–545.
- (40) Schnitker, J.; Gopalswamy, R.; Crippen, G. M.; Objective models for steroid binding sites of human globulins. *J. Comput. Aid. Mol. Des.* **1997**, *11* (1), 93–110.
- (41) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, New 3D Theoretical Descriptors Derived from Molecular Surface Properties: A Comparative 3D QSAR Study in a Series of Steroids. *J. Comput.-Aid. Mol. Des.* **1997**, *11*, 79–92.
- (42) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G.; Alignment of molecules by the Monte Carlo optimization of molecular similarity indices. *J. Comput. Chem.* **1997**, *18* (11), 1344–1353.
- (43) Klebe, G.; Abraham, U.; Mietzer, T. Molecular Similarity Indexes in a Comparative-Analysis (COMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130.
- (44) Hahn, M.; Rogers, D. Receptor Surface Models. 2. Application to Quantitative Structure–Activity Relationships Studies. *J. Med. Chem.* **1995**, *38*, 2091–2102.
- (45) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modelling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.
- (46) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure–Activity Relationships from Molecular Similarity Matrixes and Genetic Neural Networks. 1. Method and Validations. *J. Med. Chem.* **1997**, *40*, 4347–4359.
- (47) Silverman, B. D.; Plat, D. E. Comparative Molecular Moment Analysis (CoMMA): 3D-QSAR without Molecular Superposition. *J. Med. Chem.* **1996**, *39*, 2129–2140.
- (48) Krystek, S. R.; Hunt, J. T.; Stein, P. D.; Stouch, T. R. Three-Dimensional Quantitative Structure–Activity Relationships of Sulfonamide Endothelin Inhibitors. Sulfonamide Endothelin Inhibitors. *J. Med. Chem.* **1995**, *38*, 659–668.
- (49) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P.; A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- (50) Stewart, J. J. P. Mopac 6.0, Quantum Chemical Program Exchange, 455, 1990.
- (51) RATTLE, Oxford Molecular Ltd, Oxford, U.K.
- (52) Ferenczy, G.; Reynolds, C. A.; Richards, W. G. Semiempirical AM1 Electrostatic Potentials and AM1 Electrostatic Potential Derived Charges: A Comparison with Ab Initio Values. *J. Comput. Chem.* **1990**, *11*, 159.
- (53) Westphal, U. *Steroid-Protein Interactions II*; Springer-Verlag: Berlin, 1986.
- (54) Dunn, J. F.; Nisula, B. C.; Rodbard, D. Transport of Steroid Hormones: Binding of 21 Endogenous Steroids to Both Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, *53*, 58–68.
- (55) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid Protein Interactions. Human Corticosteroid Binding Globulin: Physicochemical Properties and Binding Specificity. *Biochemistry* **1981**, *20*, 6211–6218.
- (56) Pugeat, M. M.; Dunn, J. F.; Nisula, B. C. Transport of Steroid Hormones: Interaction of 70 Drugs with Testosterone-Binding Globulin and Corticosteroid-Binding Globulin in Human Plasma. *J. Clin. Endocrin. Metab.* **1981**, *53*, 69–75.
- (57) Winn, P. J.; Lyne, P. D.; Richards, W. G. Unpublished results.

JM9810607