

# The Application of Molecular Similarity Calculations

Catherine Burt and W. Graham Richards\*

Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, United Kingdom.

Philip Huxley

Ciba-Geigy AG, R-10596.09, CH 4002, Basel, Switzerland.

Received 23 November 1989; accepted 18 June 1990

A prescription for applying the method of molecular similarity calculations based on electrostatic potentials and fields is developed by consideration of a typical structure-activity series. Firm conclusions are drawn about the nature of the grid of points surrounding the molecules and about the choice of geometry, but options for point charges are less clearcut.

## INTRODUCTION

Quantitative measures of molecular similarity have great appeal in structure-activity studies. However, although several articles<sup>1-3</sup> have outlined various measures of similarity and aspects of implementation, there has been no clear prescription of just how similarity is best used and a number of questions remain about how to apply what is widely accepted to be an interesting idea. Here we try to produce a mode for the application by investigating options on a real set of data from the agrochemical field using a new molecular similarity computer program.

Carbo<sup>4,5</sup> pioneered the idea of quantitative measures of similarity introducing a formula in terms of electron density overlap between superimposed molecules *A* and *B*

$$R_{AB} = \frac{\int \rho_A \rho_B dv}{\left( \int \rho_A^2 dv \right)^{1/2} \left( \int \rho_B^2 dv \right)^{1/2}} \quad (1)$$

where  $\rho_A$  = electron density of molecule *A*.

An alternative index due to Hodgkin<sup>6</sup> compares the magnitude as well as the shape of the electron density.

$$H_{AB} = \frac{2 \int \rho_A \rho_B dv}{\int \rho_A^2 dv + \int \rho_B^2 dv} \quad (2)$$

It can be shown that the Hodgkin index will always be less than or equal to the Carbo index.<sup>7</sup> Although originally proposed as a method of com-

paring molecules in terms of electron density, the formulae can be used with other calculable parameters,<sup>8</sup> providing singularities can be avoided. The use of electrostatic potentials and electrostatic fields is particularly attractive since they are better discriminators than charge and problems can be avoided if only the values external to a van der Waals volume of the molecule are considered. Another advantage of electrostatic potential and electrostatic field is that they can rapidly be computed from point charges or other charge distribution parameters. The potentials or fields can be calculated over a grid of points surrounding a molecule. What remains unresolved is just how fine this grid of points must be and how far beyond the molecular surface the grid should extend. It should be noted that the Carbo electron density approach is not based on calculation over a grid.<sup>9</sup>

These questions are investigated here using a comprehensive new computer program. In addition the questions of which molecular geometries are safe to use and which source of atom-based charges provide the most reliable results are probed, with the intention of providing a clear prescription for applications of molecular similarity.

## THE DATA

Soloway et al.<sup>10</sup> have published an extensive study of nitromethylene insecticides including detailed biological data. Many of the compounds show extraordinary potency, insecticidal activity being measured by exposing third instar larvae of the corn-ear worm (*Heliothis Zea*) to sprayed bean plants (*Vicia Faba*) and being expressed relative to that of parathion (taken to be 100). Rajappa<sup>11</sup> has comprehensively reviewed their

\*To whom all correspondence should be addressed.

preparation, structure and synthetical potential and Odell<sup>12</sup> has performed a pharmacophoric mapping study. The structure and relative activity of the principal nitromethylene heterocycles are given in Table I.

### Data Preparation

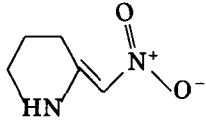
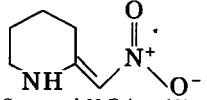
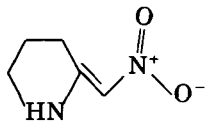
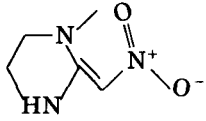
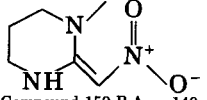
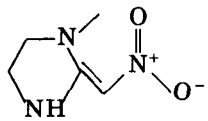
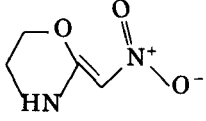
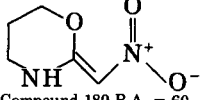
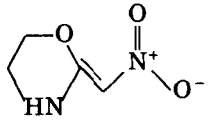
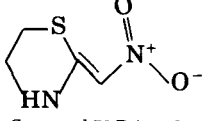
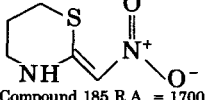
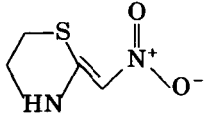
Both *E* and *Z* isomers of the molecules were modeled using the molecular modeling package MACROMODEL 2.5.<sup>13</sup> The geometries were minimized using the program's interface to the MM2<sup>14</sup> program. These "MM2" geometries were used as the starting point for geometry optimizations using the program MOPAC (version 5.0)<sup>15</sup> with the AM1 Hamiltonian.<sup>16</sup> The two geometries will henceforth be referred to as "MM2" and "AM1." Atomic charges were calculated for both sets of geometries using AM1, MNDO<sup>17</sup> (i.e., MOPAC 5.0) and STO-3G (GAMESS program<sup>18</sup>) methods. Both Mulliken and Löwdin charges were extracted from the GAMESS output files. The following conventions will henceforth be used to describe the charges, "AM1," "MNDO," "MUL," and "LOW." A further convention used in naming the structures concerns isomerism about the double bond. We are concerned with the relative positions of the nitro and amino groups. "O" structures always have the nitro and amino groups on opposite sides of the double bond, whereas "S" structures always have these groups on the same side of the double bond. This is to facilitate file handling while at the same time preventing a contradiction of the IUPAC convention of *E/Z* nomenclature.

In order for the molecular similarity indices to be meaningful, the molecules must be oriented in a "reasonable" position relative to one another. Since compound 185 has the highest relative activity, it was taken to be the "lead" compound with which all other molecules are compared in calculating similarity indices. Molecule 185 was, therefore, oriented with the double bond lying along the *x* axis. The remaining compounds were "fitted" to compound 185 within MACROMODEL 2.5 by performing a least squares fitting of the following corresponding atoms of the molecules, the N of the nitro group, the N of the amino group and the two C atoms of the double bond.

### Calculation of Molecular Similarity Indices

The automatic similarity package, ASP, constructs a grid around each pair of molecules and computes the electrostatic potential and electric field at certain points on the grid. The default calculation of the program excludes all grid points lying within molecular volumes from the overlap integral of the numerator and the relevant term of the denominator<sup>6</sup> of the molecular similarity indices. This is to prevent the occurrence of spurious singularities. Molecular similarity indices may be optimized and the program employs the SIMPLEX<sup>19</sup> method. The lead compound is fixed and the molecule whose similarity index is to be optimized is rotated and translated in space until it lies in its position of maximum similarity with respect to the fixed molecule.

**Table I.** The structure and relative activity of the principal nitromethylene heterocycles.

 Compound 72 R.A. = 90	 Compound 29 R.A. = 160	 Compound 85 R.A. = 10
 Compound 113 R.A. = 300	 Compound 150 R.A. = 140	 Compound 161 R.A. = 30
 Compound 168 R.A. = 0	 Compound 180 R.A. = 60	 Compound 999 R.A. = unmeasured
 Compound 7 R.A. = 8	 Compound 185 R.A. = 1700	 Compound 300 R.A. = 40

In order to analyze similarity indices, we need first to know the optimum grid which must be used in the computation of the indices. There are two relevant grid parameters, the grid extent, i.e., the distance the grid extends from the lowest and highest  $x$ ,  $y$ , and  $z$  coordinates of the atoms of the molecules and the grid increment, i.e., the spacing between the lines of the grid. Needless to say, the coarser the grid, the quicker the calculation but this must be balanced by loss of accuracy.

Molecular similarity indices were calculated (using AM1 geometries and AM1 charges) for various grid extents (4, 6, 8, 10, and 20 Å) and grid increments (1 and 2 Å). The Z isomer of compound 185, which is named 185O according to my convention was taken to be the lead compound. E and Z isomers (O and S structures) of each compound were compared with the lead.

Correlation of the grid parameters during optimization was initially investigated by optimizing Hodgkin Electric Field Molecular Similarity (EFMS) indices for compounds 72O and 72S with respect to compound 185O, using the different grids above.

In order to investigate whether a coarser grid would produce maxima corresponding to those produced using a finer grid, for the full data set of compounds, Hodgkin EFMS indices were optimized using a 4 Å/1 Å grid. The same data set was optimized using a 4 Å/2 Å grid and the corresponding values of the molecular similarity index of each of the molecules in their optimum positions were calculated using the 4 Å/1 Å grid and compared with their initial values.

Once the optimum grid parameters had been chosen, molecular similarity indices were calculated and optimized for both geometries with all four charge schemes (i.e., eight schemes in all). The correlation between the different charge schemes and the different geometries was investigated using the LOTUS<sup>20</sup> package.

Finally, the correlation of the various similarity indices with biological activity was investigated.

## RESULTS AND DISCUSSION

Table II shows the variation in C.P.U. time with grid parameters for the series of single point (i.e., nonoptimizing) molecular similarity calculations using AM1 geometries and AM1 charge schemes. It can be seen that the "coarser" grids with a grid increment of 2 Å show appreciable saving in C.P.U. time. The values of all the similarity indices for each grid are too numerous to list here, but some are given in Tables III and IV. The similarity indices for all the different schemes were plotted against compound number. Here, it is the shapes of the curves that is important and this seemed to depend solely upon grid increment, the coarser 2 Å grid increment giving a different ranking of the similarity indices than the more accurate 1 Å grid increment. Ranking of the similarity indices did not appear to depend on grid extent, the 4 Å/1 Å curve "tracking" the 20 Å/1 Å curve. Hence the 4 Å/1 Å curve was deemed to be best.

Table V shows the variation in C.P.U. time when the Hodgkin EFMS indices of compounds 72O and 72S are optimised with respect to compound 185O for the different sets of grid parameters. Again it can be seen that the coarser 2 Å grid gives an appreciable reduction in C.P.U. time.

Table VI gives a comparison of the difference between final and initial values ( $\Delta$ ) when the Hodgkin EFMS index is optimized for compounds 72O and 72S and also shows the number of iterations of the SIMPLEX loop for the varying grid parameters. When a 1 Å grid increment is used, for compound 72O,  $\Delta$  always takes the value of 0.026 and optimization results from 47 iterations of the SIMPLEX loop, regardless of the grid extent. For compound 72S,  $\Delta$  varies by a mere 0.006, a 4 Å taking 61 iterations of the SIMPLEX loop to optimize, all other grid extents taking 53 iterations. For a 2 Å grid increment, for compound 72O the 4 Å grid extent maximum differs from all the others taking 63 iterations of the SIMPLEX loop to produce a  $\Delta$  value of 0.062. All other grid

**Table II.** Variation of C.P.U. time with grid parameters for single point calculations on full data set.

Grid Extent/Å	Grid Increment/Å	C.P.U. Time/hh:mm:ss.ss
4	1	00:04:10.22
4	2	00:00:38.16
6	1	00:07:42.30
6	2	00:01:07.68
8	1	00:13:01.92
8	2	00:01:51.52
10	1	00:19:59.86
10	2	00:02:44.90
20	1	01:35:01.96
20	2	00:12:31.40

**Table III.** Single-point Carbo EPMS indices for various grid parameters.

Compound	4 Å/1 Å value	4 Å/2 Å value	20 Å/1 Å value	20 Å/2 Å value
185S	0.826	0.837	0.858	0.865
72O	0.949	0.960	0.959	0.967
72S	0.830	0.801	0.858	0.837
113O	0.942	0.955	0.956	0.965
113S	0.827	0.833	0.854	0.858
168O	0.922	0.953	0.938	0.961
168S	0.855	0.882	0.888	0.908
172O	0.970	0.972	0.978	0.979
172S	0.816	0.816	0.847	0.846
29O	0.976	0.966	0.981	0.974
29S	0.839	0.846	0.868	0.873
150O	0.970	0.980	0.976	0.983
150S	0.817	0.806	0.847	0.838
180O	0.963	0.961	0.972	0.969
180S	0.872	0.878	0.900	0.905
85O	0.960	0.945	0.970	0.958
85S	0.824	0.814	0.855	0.848
161O	0.933	0.939	0.949	0.953
161S	0.812	0.835	0.843	0.860
999O	0.919	0.935	0.933	0.945
999S	0.831	0.864	0.864	0.888
300O	0.939	0.917	0.955	0.937
300S	0.809	0.826	0.843	0.855

**Table IV.** Single-point Hodgkin EFMS indices for various grid parameters.

Compound	4 Å/1 Å value	4 Å/2 Å value	20 Å/2 Å value	20 Å/2 Å value
185S	0.567	0.558	0.581	0.570
72O	0.885	0.900	0.888	0.903
72S	0.585	0.555	0.597	0.567
113O	0.818	0.820	0.824	0.826
113S	0.623	0.630	0.634	0.640
168O	0.795	0.833	0.801	0.838
168S	0.548	0.590	0.563	0.602
172O	0.933	0.865	0.935	0.870
172S	0.563	0.583	0.576	0.594
29O	0.933	0.919	0.935	0.922
29S	0.616	0.584	0.627	0.595
150O	0.894	0.879	0.898	0.883
150S	0.573	0.578	0.586	0.589
180O	0.879	0.889	0.883	0.892
180S	0.577	0.543	0.591	0.556
85O	0.911	0.892	0.914	0.895
85S	0.574	0.516	0.587	0.527
161O	0.769	0.781	0.776	0.787
161S	0.573	0.591	0.585	0.601
999O	0.804	0.794	0.810	0.800
999S	0.542	0.608	0.556	0.618
300O	0.801	0.731	0.808	0.738
300S	0.544	0.539	0.558	0.550

extents take 48 iterations of the SIMPLEX to produce  $\Delta \approx 0.058$ . Compound 72S tended to converge in 47 iterations with  $\Delta \approx 0.218$ . However, in this case the 4 Å/2 Å grid (number of iterations = 83,  $\Delta = 0.251$ ) and the 20 Å/2 Å (number of iterations = 68,  $\Delta = 0.245$ ) were exceptions. The overall results again seem to indicate that grid extent has little effect on the relative values of the molecular similarity indices.

Similarity indices are not exact parameters and so the question arises as to whether it is better to (1) use the 4 Å/1 Å grid for an initial single point, nonoptimizing calculation, optimize using a 4 Å/2 Å and then recalculate the similarity indices of the molecule in their "optimum" positions using the 4 Å/1 Å grid. The assumption here is that the 4 Å/2 Å grid maximum would be close to the 4 Å/1 Å grid maximum and the sav-

**Table V.** The variation in the C.P.U. time when the Hodgkin EFMS indices of compounds 72O and 72S are optimized with respect to compound 185O, with grid parameters.

Grid Extent (Å)	Grid Increment (Å)	C.P.U. Time (hh:m:ss.ss)
4	1	00:41:43.19
4	2	00:06:46.76
6	1	01:12:13.48
6	2	00:08:44.07
8	1	01:58:46.89
8	2	00:14:35.46
10	1	13:00:34.59
10	2	00:22:14.23
20	1	14:30:26.08
20	2	01:52:45.72

**Table VI.** Survey of the effect of grid parameter on optimization of the Hodgkin EFMS index.

Grid ext.	Grid inc.	Compound	Initial value	Final value	$\Delta$	No. of increments
4	1	72O	0.885	0.911	0.026	47
		72S	0.585	0.763	0.178	61
4	2	72O	0.900	0.962	0.062	63
		72S	0.555	0.806	0.251	83
6	1	72O	0.887	0.913	0.026	47
		72S	0.591	0.765	0.174	53
6	2	72O	0.901	0.960	0.059	48
		72S	0.561	0.781	0.220	47
8	1	72O	0.887	0.913	0.026	47
		72S	0.593	0.767	0.174	53
8	2	72O	0.902	0.960	0.058	48
		72S	0.564	0.782	0.218	47
10	1	72O	0.888	0.914	0.026	47
		72S	0.595	0.768	0.173	53
10	2	72O	0.902	0.960	0.058	48
		72S	0.565	0.783	0.218	47
20	1	72O	0.888	0.914	0.026	47
		72S	0.597	0.769	0.172	53
20	2	72O	0.903	0.961	0.058	48
		72S	0.567	0.812	0.245	68

ing in C.P.U. time would be considerable. A final set of 4 Å/1 Å grid single point calculations are necessary since similarity indices calculated with different grid parameters are not directly comparable; or (2) optimize using the 4 Å/1 Å grid. In this case there will be a dramatic increase in the C.P.U. time but the optimized indices will be directly comparable with the initial values.

Table VII summarizes the results of the above calculations. The results quite categorically show that optimized values of "single point" indices produced by a given set of grid parameters should not be produced with a different set of parameters since the two series are not directly comparable. Indeed the worst possible outcome can occur in which the 4 Å/2 Å optimized-4 Å/1 Å single-point value (fifth column, Table VII) is less than the initial 4 Å/1 Å single point value (column 2, Table VII). This indeed was the case 17 out of the

23 cases above. For example, consider compound 113O

4 Å/1 Å single-point value of Hodgkin EFMS index = 0.818

4 Å/1 Å optimized value = 0.839

4 Å/2 Å optimized value = 0.864

4 Å/1 Å single-point value for 4 Å/2 Å optimized position = 0.807 and 0.807 < 0.818.

the 4 Å/1 Å optimization took 08:51:26.28 of C.P.U. time. The 4 Å/2 Å optimization took 01:11:11.96 of C.P.U. time followed by the 4 Å/1 Å single-point calculation at 03:59.84 of C.P.U. time, giving a total of 01:15:11.80. The increase in C.P.U. time is significant but not prohibitive.

Hence it is concluded that the ranking of the molecular similarity indices does not vary with grid extent, but does vary with grid increment in

**Table VII.** Comparison of optimized values of the Hodgkin EFMS index for varying grid parameters.

Compound	Initial value 4 Å/1 Å grid	Optimized value 4 Å/1 Å grid	Optimized value 4 Å/1 Å grid	Single-point value of 4 Å/2 Å optimized index using 4 Å/1 Å grid
185S	0.567	0.758	0.823	0.720
72O	0.885	0.911	0.962	0.614
72S	0.585	0.763	0.806	0.540
113O	0.818	0.839	0.864	0.807
113S	0.623	0.781	0.805	0.722
168O	0.795	0.848	0.902	0.789
168S	0.548	0.660	0.729	0.645
172O	0.933	0.948	0.974	0.872
172S	0.563	0.734	0.776	0.691
29O	0.933	0.948	0.974	0.699
29S	0.616	0.776	0.825	0.563
150O	0.894	0.900	0.930	0.752
150S	0.573	0.784	0.799	0.562
180O	0.879	0.893	0.946	0.856
180S	0.577	0.705	0.726	0.634
85O	0.911	0.932	0.958	0.640
85S	0.574	0.743	0.746	0.502
161O	0.769	0.803	0.837	0.519
161S	0.573	0.737	0.780	0.416
999O	0.804	0.841	0.841	0.642
999S	0.542	0.692	0.710	0.455
300O	0.801	0.820	0.838	0.791
300S	0.544	0.740	0.705	0.656

the range studied, a 2 Å grid being too coarse. The optimum grid is, therefore, a 4 Å/1 Å grid for both single-point and optimization calculations. All further calculations were performed using this grid.

### Correlation of Different Charge Schemes

For the O structures fitted to compound 185O and for each given geometry and similarity index, graphs were drawn of the similarity indices produced by one charge scheme against those produced by another charge scheme and regression analysis was performed using the LOTUS package. In this way each charge scheme was compared with every other one for the eight permutations of geometry-similarity index. The *R* squared values are given in Tables VIII. As expected the AM1 and MNDO charges correlate well with each other, as do STO-3G Mulliken charges with STO-3G Löwdin charges. Correlation between AM1 and STO-3G Löwdin charges and between MNDO and STO-3G Löwdin charges is also good. Correlation between AM1 and STO-3G Mulliken charges is less good and correlation between MNDO and STO-3G Mulliken charges is sometimes quite poor. For AM1 geometries, where correlation is poor, it is found that compound 168 and 180 are usually responsible. For MM2 geometries, compounds 180 and 999 tend to give poor correlation. These are all compounds with an O atom at position three with respect to the amino nitrogen.

### Correlation of Different Geometries

For the O structures fitted to compound 185O and for each charge scheme and similarity index, the similarity indices produced by the MM2 geometry scheme were plotted against those produced by the AM1 geometry scheme. The *R* squared values are given in Tables IX. Overall, the different geometries correlate very badly with one another. Compounds 300, 168, 150, 113, and 999 give a particularly poor correlation. With the exclusion of compound 168, this poor correlation can be explained by the large difference in the dihedral angle defined by the C atoms of the double bond, the N and an O of the nitro group (Table X).

### Correlation with Biological Activity

For each geometry and charge scheme, plots of each similarity index versus biological activity were produced for the O structures fitted to compound 185O and for the S structures fitted to compound 185S. Similar plots were obtained for the optimized values of these indices. There are far too many plots to show here. The most striking result is shown in Figure 1 in which the Hodgkin EPMS indices are plotted against biological activity for the O structures fitted to compound 185O and for the AM1 geometry/AM1 charge scheme. There are 10 points on the curve and a perfect ranking is given in all but three cases. Compound 113 is predicted to have a lower

**Table VIII.** Correlation of the different charge schemes used to calculate molecular similarity indices.

	AM1	MNDO	MVL
AM1 geometry, CARBO EPMS index			
MNDO	0.960		
MUL	0.757	0.626	
LOW	0.887	0.806	0.903
AM1 geometry, HODGKIN EPMS index			
MNDO	0.958		
MUL	0.630	0.483	
LOW	0.835	0.727	0.859
AM1 geometry, CARBO EFMS index			
MNDO	0.984		
MUL	0.921	0.872	
LOW	0.950	0.942	0.927
AM1 geometry, HODGKIN EFMS index			
MNDO	0.984		
MUL	0.929	0.876	
LOW	0.952	0.942	0.924
MM2 geometry, CARBO EPMS index			
MNDO	0.956		
MUL	0.659	0.486	
LOW	0.720	0.558	0.987
MM2 geometry, HODGKIN EPMS index			
MNDO	0.971		
MUL	0.638	0.506	
LOW	0.762	0.648	0.957
MM2 geometry, CARBO EFMS index			
MNDO	0.983		
MUL	0.850	0.788	
LOW	0.904	0.850	0.981
MM2 geometry, HODGKIN EFMS index			
MNDO	0.986		
MUL	0.875	0.828	
LOW	0.918	0.875	0.981
Average			
MNDO	0.973		
MUL	0.782	0.683	
LOW	0.866	0.794	0.940

activity and compounds 172 and 85 higher activity. Prediction of too large an activity may not be serious as the biological activity may be reduced by some extraneous factor but a low prediction could lead to the neglect of a useful molecule. Only one instance of the latter problem occurs.

In the corresponding plots of the Carbo EPMS indices (not shown) compound 180 is raised above the line and, therefore, predicted to have a higher similarity and activity. This compound has an O atom in position three with respect to the amino nitrogen and in the same position as the S atom of the lead compound. This would seem to vindicate Hodgkin's method of normalization, which takes into account magnitude as well as the shape of the charge distribution.

Optimization of these indices did not alter the ranking or the shape of the curves appreciably. This seems to indicate that the fitting to the lead compound was correct and that perhaps the nitroamine group is held with the nitro and amino groups trans to each other in a specific orientation/position at the binding site of the receptor. For

**Table IX.** Correlation of different geometry schemes used to calculate molecular similarity indices.

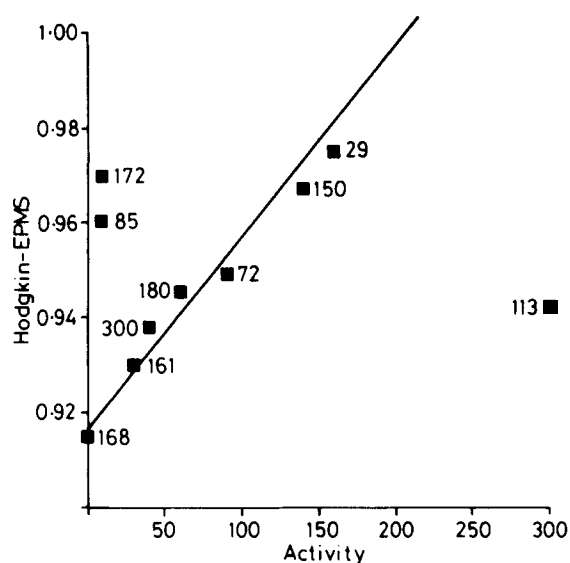
	AM1	MM2
a	AM1 charges, CARBO EPMS index	0.137
b	AM1 charges, HODGKIN EPMS index	0.187
c	AM1 charges, CARBO EFMS index	0.084
d	AM1 charges, HODGKIN EFMS index	0.102
e	MNDO charges, CARBO EPMS index	0.173
f	MNDO charges, HODGKIN EPMS index	0.246
g	MNDO charges, CARBO EFMS index	0.043
h	MNDO charges, HODGKIN EFMS index	0.063
i	MUL charges, CARBO EPMS index	0.041
j	MUL charges, HODGKIN EPMS index	0.023
k	MUL charges, CARBO EFMS index	0.066
l	MUL charges, HODGKIN EFMS index	0.062
m	LOW charges, CARBO EPMS index	0.008
n	LOW charges, HODGKIN EPMS index	0.003
o	LOW charges, CARBO EFMS index	0.001
p	LOW charges, HODGKIN EFMS index	0.002
q	Average	MM2
		0.078

**Table X.** Dihedral angles defined by the C atoms of the double bond, the N and an O atom of the nitro group for O structures for AM1 and MM2 geometries.

Compound	Dihedral angle		Difference
	AM1	MM2	
168	-1.3	0.0	1.3
300	-32.1	-5.1	27.0
150	2.1	-29.7	31.8
113	-6.9	17.0	23.9
999	-15.6	-1.1	14.5

the AM1 geometry scheme, the MNDO charge scheme gives the same ranking as the AM1 charge scheme but STO-3G Mulliken and STO-3G Löwdin charge schemes seem to overestimate the oxygen containing compounds 168 and 180 (compound 999 is not included since its activity is unmeasured). The MM2 geometry scheme does not give as good a correlation with biological activity. Correlation of the similarity indices of S structures fitted to compound 185S with biological activity is also less clear.

Hence it is concluded that the AM1 geometries together with AM1 charges produced by the pro-



**Figure 1.** Plot of the Hodgkin Electrostatic Potential Molecular Similarity indices versus biological activity for the O structures relative to compound 185O for the AM1 geometry/AM1 charge scheme.

gram MOPAC are the best of the methods tried in calculating similarity indices for this class of compounds.

## CONCLUSION

The purpose of this article was to provide a prescription as to how molecular similarity calculations may best be applied. We have found that a grid with a 4 Å extent and 1 Å increment was the optimum one to use in the integration of the indices. The Hodgkin index gave a better measure of the similarity than the Carbo index, most notably when comparing oxygen containing compounds to the lead, sulphur containing compound. AM1 geometries were found to be superior to MM2 geometries and together with the AM1 charge scheme, gave a reasonable correlation with biological activity. These AM1 charges are, however, produced using Mulliken population analysis,<sup>21</sup> in which the charge between two nuclei in the molecule is divided equally between the two, regardless of the electronegativities of the atoms. More accurate charges would be obtained by fitting point charges to reproduce electron densities calculated from the *ab initio* wave function and there are Q.C.P.E. programs<sup>22</sup> which will generate these point charges. However, this would dramatically increase the amount of C.P.U. time needed to perform the calculations. Ferenczy et al.<sup>23</sup> have developed a semiempirical method of calculating atomic charges derived from AM1 potentials of comparable quality to STO-3G potential derived charges. Similarity indices calculated using these charges may give an even better correlation with biological activity.

We have provided a new computer program<sup>24</sup> that rapidly calculates molecular similarity indices *reproducibly* and is easily applicable to a bioactive series of compounds, containing several heavy atoms. With the molecular similarity method now in a package form, it is hoped that the technique will become a major weapon in the armoury of the medicinal chemist.

These calculations were carried out using the computing facilities at Ciba-Geigy AG, Basel. The authors also thank Drs. C.M. Edge and R.M. Hindley of Beecham Pharmaceuticals for helpful discussions.

## References

1. P. E. Bowen-Jenkins, D. L. Cooper, and W. G. Richards, *J. Phys. Chem.*, **89**, 2195 (1985).
2. P. E. Bowen-Jenkins and W. G. Richards, *Int. J. Quant. Chem.*, **30**, 763 (1986).
3. E. E. Hodgkin and W. G. Richards, *J. Chem. Soc., Chem. Comm.*, **1342** (1986).
4. R. Carbo, L. Leyda, and M. Arnau, *Int. J. Quant. Chem.*, **17**, 1185 (1980).
5. R. Carbo and L. Domingo, *Int. J. Quant. Chem.*, **32**, 517 (1987).
6. E. E. Hodgkin and W. G. Richards, *Int. J. Quant. Chem.*, **14**, 105 (1987).
7. A. Meyer and W. G. Richards, *J. Chem. Soc. Perkin II* (in press).
8. R. Carbo, E. Sune, F. Lapena, and J. Perez, *J. Biol. Phys.*, **14**, 21 (1986).
9. R. Carbo and B. Calabuig, *Computer Phys. Commun.*, **55**, 117 (1989).
10. S. B. Soloway, A. C. Henry, W. D. Kollmeyer, W. M. Padgett, J. E. Powell, S. A. Roman, C. H. Tiemann, R. A. Corey, and C. A. Horne, *Pestic. Venom. Neurotoxic.*, [Sel. Pap. Int. Congr. Entomol.], Ed. D. L. Shankland, R. M. Hollingworth and T. Smyth, Plenum: New York, 1978, p. 153.
11. S. Rajappa, *Tetrahedron*, **37**, 1453 (1981).
12. B. Odell, *Journal of Computer Aided Molecular Design*, **2**, 191 (1988).
13. MacroModel 2.5, Clark Still, University of Columbia.
14. N. L. Alinger, *J. Amer. Chem. Soc.*, **99**, 8127 (1977).
15. J. J. P. Stewart, Q.C.P.E. 455, MOPAC 5.0.
16. M. J. S. Dewar, E. G. Zoebisch, E. P. Healy, and J. J. P. Stewart, *J. Am. Chem. Soc.*, **107**, 3902 (1985).
17. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, **99**, 4899 (1977).
18. GAMESS, M. W. Schmidt, J. A. Boatz, K. K. Baldridge, S. Koseki, M. S. Gordon, S. T. Elbert, B. Lam, *Q.C.P.E. Bulletin*, **7**, 115 (1987).
19. J. A. Nelder and R. Mead, *Comput. J.*, **7**, 308 (1965).
20. LOTUS-1-2-3, Release 2.01, Lotus Development Corporation, 55 Cambridge Parkway, Cambridge, MA 02142 (1988).
21. R. S. Mulliken, *J. Phys. Chem.*, **23**, 1833 (1955).
22. L. E. Chirlian and M. M. Francl, *Q.C.P.E. Bull.*, **7**, 39 (1987), Q.C.P.E. 524; CHELP.
23. G. G. Ferenczy, C. A. Reynolds, and W. G. Richards, *Computational Chemistry*, **11**, 159 (1990).
24. ASP-Automatic Similarity Package, Oxford Molecular, Terrapin House, South Parks Road, Oxford OX1 3UB, U.K.