Rapid Evaluation of Shape Similarity Using Gaussian Functions

A. C. Good and W. G. Richards*

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

Received April 27, 1992

An analytic technique for the comparison of molecular shape is presented. The new procedure fits Gaussian functions to the STO3G atomic orbital derived electron density functions of different atom types. The Gaussian functions are then used analytically within the Carbo similarity index. Similarity results produced by these functions are evaluated rapidly (2 or 3 orders of magnitude faster than the previous grid-based evaluation technique), greatly enhancing the potential flexibility of these calculations.

INTRODUCTION

Molecular similarity calculations are now being widely applied in molecular modeling. A number of evaluation methods have been presented,[1-16] one of the major techniques being the Carbo similarity index:[1-9]

$$R_{\mathrm{AB}} = \frac{\int P_{\mathrm{A}} P_{\mathrm{B}}\,\mathrm{d}\nu}{(\int P_{\mathrm{A}}^2\,\mathrm{d}\nu)^{1/2}(\int P_{\mathrm{B}}^2\,\mathrm{d}\nu)^{1/2}}$$

Molecular similarity R_{AB} is determined from the structural properties P_{A} and P_{B} of the two molecules being compared. The numerator measures property overlap, while the denominator normalizes the similarity result. As originally applied by Carbo,[1,2] quantum mechanically derived electron density is used as the structural property P. The technique has since been extended to cover electrostatic potentials and electric fields.[3-8]

More recently, Meyer[9] modified the index to permit the evaluation of molecular shape similarity. The mechanics of similarity evaluation are the same as those applied to electrostatic potential and electric field calculations. The molecules are surrounded by a rectilinear grid, and the structural property is evaluated at each intersection. For shape, every grid point is tested to see whether it falls inside the van der Waals surface of each molecule. The results are then applied to the following equation:

$$S_{\mathrm{AB}} = B/(T_{\mathrm{A}} T_{\mathrm{B}})^{1/2}$$

B is the number of grid points falling inside both molecules, while T_{A} and T_{B} are the total number of grid points falling inside each individual molecule.

Grid-based shape and electrostatic potential similarity evaluations, while faster than the original quantum mechanically based calculations, are still time-consuming processes. Indeed, shape calculations are slower than those for electrostatic potential, since very fine grids (0.2-Å separation is generally used) are required to obtain precise results.

Recently, work has been carried out using Gaussian functions to speed up the evaluation of electrostatic potential similarity.[17] Studies have also been undertaken fitting Gaussians to molecular fragment electron densities.[18] In this paper, we extend these ideas to embrace the calculation of shape similarity.

GAUSSIAN FUNCTION APPROXIMATION OF ATOMIC ORBITAL ELECTRON DENSITY

Instead of using van der Waals radius data, atomic electron density functions are used to describe the shape of each atom. These functions are determined from the square of the STO3G atomic orbital wave functions.[19] Three Gaussian functions were then fitted to the resulting electron density of each atom type.[20] Rather than using a grid to determine shape similarity, these Gaussian functions are then used analytically within the Carbo formula:

$$R_{\mathrm{AB}} = [\sum_{i=1}^{n}\sum_{j=1}^{m}\int (G_a^i + G_b^i + G_c^i)(G_x^j + G_y^j + G_z^j)\,\mathrm{d}\nu]/ \\ [(\sum_{i=1}^{n}\sum_{i=1}^{n}(\int (G_a^i + G_b^i + G_c^i)^2\,\mathrm{d}\nu)^{1/2}) \times \\ (\sum_{j=1}^{m}\sum_{j=1}^{m}(\int (G_x^j + G_y^j + G_z^j)^2\,\mathrm{d}\nu)^{1/2})] \quad (1)$$

where $G_z^j = \gamma_z e^{-\alpha_z(r-R_j)^2}$ and R_j is the nuclear coordinate position of atom j.

Equation 1 expands into a series of two center Gaussian overlap integrals. The two center integrals have a simple form made up of exponent values and atom center distances.[21] For example:

$$\int e^{-\alpha_1(r-R_i)^2} e^{-\alpha_2(r-R_j)^2}\,\mathrm{d}\nu = \left(\frac{\pi}{\alpha_1 + \alpha_2}\right)^{3/2} \times \\ \exp\left(\frac{\alpha_1\alpha_2}{\alpha_1 + \alpha_2}|R_i - R_j|^2\right) \quad (2)$$

It is thus possible to determine shape similarity through a series of readily calculable exponent terms.

Three ideas of atomic size were considered when calculating the Gaussian functions:

(a) Three Gaussians were fitted directly to the STO3G atomic orbital derived electron density of each atom type (henceforth known as "unmodified Gaussians").

(b) To account for atoms appearing harder within molecules, the three Gaussians were also fitted to a modified version of the electron density function (henceforth referred to as "VDW fitted Gaussians"). For this modified function, the electron density is set to zero beyond the van der Waals radius of each atom type.

(c) Hydrogen atoms are particularly soft, i.e., electron density extends well beyond the van der Waals radius. As a consequence, it was felt that the use of the modified Gaussians might only be required for the hydrogen atom type. A third set of functions was therefore created using the VDW fitted Gaussians for hydrogen and the unmodified Gaussians for all other atom types (henceforth known as "H VDW fitted Gaussians").

0095-2338/93/1633-0112$04.00/0

Table I. Exponent and Proportionality Constants for Gaussian Function Electron Density Approximations

atom type/electron density function	proportionality constant 1	proportionality constant 2	proportionality constant 3	exponent constant 1	exponent constant 2	exponent constant 3
C unmodified	1.6395	0.7755	1.0301	0.6430	2.3759	23.3845
H unmodified	0.3089	0.6283	0.3492	4.1596	1.0890	0.4206
N unmodified	0.9214	1.0466	1.4811	2.3592	27.9657	0.7678
O[a] unmodified	2.1309	1.1783		1.2864	21.6476	
S unmodified	1.7919	2.1151	0.9294	8.6866	0.5805	46.5772
C VDW fitted	1.3231	2.7598	-0.6355	15.7112	0.6112	0.2213
H VDW fitted	-36.0429	18.8820	18.5006	3.4781	2.7741	4.1394
N VDW fitted	0.7548	2.2254	0.5550	14.6074	1.0640	58.4097
O VDW fitted	0.5426	2.1365	0.7731	11.6386	1.3322	48.3193
S VDW fitted	1.9980	2.2907	0.9409	11.8773	0.6699	121.9395

[a] Two of the calculated exponent constants are identical and, therefore, combined.

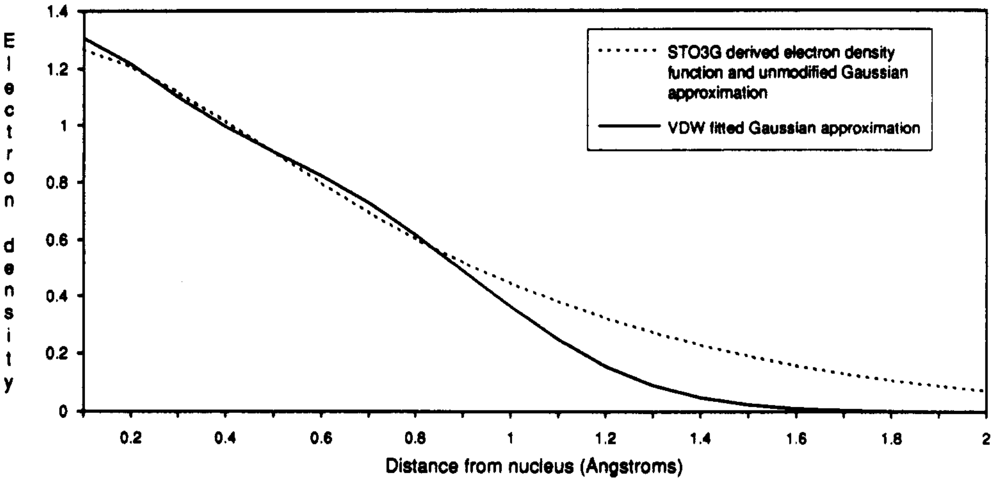
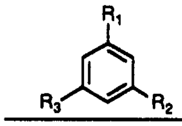


Figure 1. STO3G-derived and Gaussian approximation plots of electron density for hydrogen.

Table II. Benzene Analogues Used in Study 1

	R_1	R_2	R_3
A1	NO_2	NO_2	NO_2
A2	CH_3	CH_3	CH_3
A3	SO_2CH_3	H	H
A4	NO_2	NO_2	H
A5	SCN	H	H
A6	CH_3	CH_3	H
A7	OCH_3	H	H
A8	NO_2	H	H
A9	CH_3	H	H
A10	OH	H	H

The electron density Gaussian function approximations have been evaluated for carbon, hydrogen, oxygen, nitrogen, and sulfur. The resultant unmodified and VDW fitted Gaussian exponent and proportionality constants are given in Table I.

The electron density function, and unmodified and VDW fitted Gaussian function curves for hydrogen are shown in Figure 1. Note that the electron density and unmodified Gaussian function curves are virtually identical.

The Gaussian function overlap integrals have been evaluated for every possible combination of atom types using eq 2. The resultant exponent terms are applied to the Carbo formula in subroutines which replace the grid-based evaluation routines of the ASP program.[7] This subroutine executes a single point similarity calculation for a given set of coordinates from two candidate molecules. If similarity optimization is required, a separate routine calls this subroutine after altering the coordinates of the mobile molecule. The sign of the resultant similarity value is inverted and used as the variable to be minimized via the simplex method.[22]

ADDITIONAL SOFTWARE FEATURES

In addition to the similarity evaluation routine alterations described above, two additional software features have been tested for their impact on calculation speed and similarity result. The first is the addition of a distance cutoff option. By setting a cutoff shape overlap, calculations between atoms whose centers are separated by a distance greater than the cutoff value are not undertaken. For the second feature, all exponent terms with a proportionality constant less than 0.1 are excluded from the overlap calculation routines. This excludes approximately 40% of the exponents from the calculation. The idea behind both of these functions is to reduce the number of calculations required to evaluate shape similarity, thus increasing the speed of computation.

SIMILARITY CALCULATIONS

To compare the behavior of the new shape similarity routines with the earlier grid-based routines, four separate studies were undertaken:

(1) Benzene was compared with 10 assorted benzene analogues.

(2) Benzene was compared with 10 randomly oriented conformations of a benzene analogue.

(3) Histamine was compared with nine structures obtained from a 3-D database search of the Chapman and Hall Dictionary of Drugs.[23,24]

(4) Np-apomorphine was compared with a second Np-apomorphine molecule systematically rotated about the z axis through its nitrogen atom.

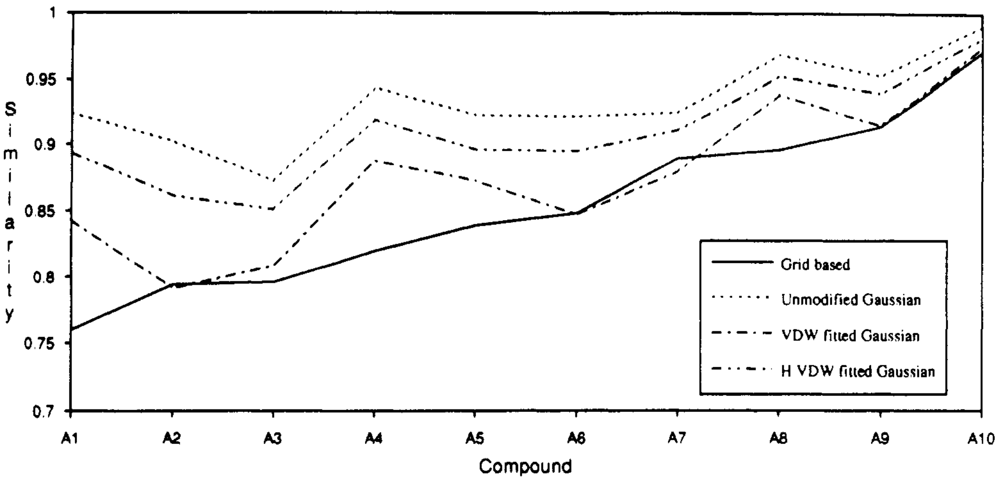


Figure 2. Similarity results for study 1.

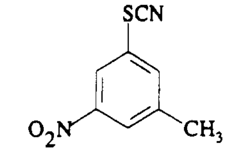


Figure 3. Benzene analogue used in study 3.

For each study, similarity calculations were executed for the unmodified, VDW fitted, and H VDW fitted Gaussian functions, together with the original grid-based evaluation method. All grid-based similarity calculations were undertaken using a 0.2-Å grid separation. The results produced in studies 1–3 (Figures 2, 4, and 7) were positioned in order of increasing grid-based similarity values.

A fifth investigation was also carried out to determine the behavior of various Gaussian function options when used in conjunction with the additional software features during similarity optimization. For this investigation, Np-apomorphine was optimized against another Np-apomorphine molecule occupying a different region of Cartesian space.

For all the studies, structures not obtained from the 3-D database were built and minimized within the Chem-X modeling program[24] using the default structural parameters and force field. The structures obtained from the 3-D database search w2ere also minimized within Chem-X using the default force field.

Study 1. The analogues were superimposed onto benzene by a least-squares fit of the ring carbon atoms. Substitutions present on each analogue are shown in Table II. The similarity results obtained for the system are displayed in Figure 2.

Study 2. The analogue was superimposed onto benzene by a least-squares fit of the ring carbon atoms. Random rotations (–40 to 90°) and translations (–2.0 to 4.0 Å) were then applied systematically about each axis through the analogue's sulfur atom. This was repeated until 10 randomly oriented conformations of the analogue were produced. The benzene analogue used is displayed in Figure 3. Similarity results obtained for the investigation are shown in Figure 4.

Study 3. Nine structures were extracted from the answer set obtained by a search of the Chapman and Hall Dictionary of Drugs[23,24] database using Chem-DBS3D.[24] The query used for the search is shown in Figure 5. The structures extracted are shown in Figure 6. The structures were superimposed onto histamine by a least-squares fit of the nitrogen atoms numbered in Figure 6. The similarity calculation results for the study are displayed in Figure 7.

A number of additional evaluations were also undertaken to test the extra software features. For the first set of calculations, a number of differing distance cutoffs (2.5–3.6 Å) were used in conjunction with the unmodified Gaussian function exponents. Similarity values obtained using a 3.6-Å cutoff (twice the van der Waals radius of the largest atom type considered) were virtually the same as those produced using no cutoff (largest difference 0.003). Differences increased as the cutoff distance was reduced, although even at 2.5 Å the results obtained were still broadly similar. For the second set of calculations, the small exponent terms were removed from the shape overlap calculations. The resulting similarity values never varied by more than 0.002 from those

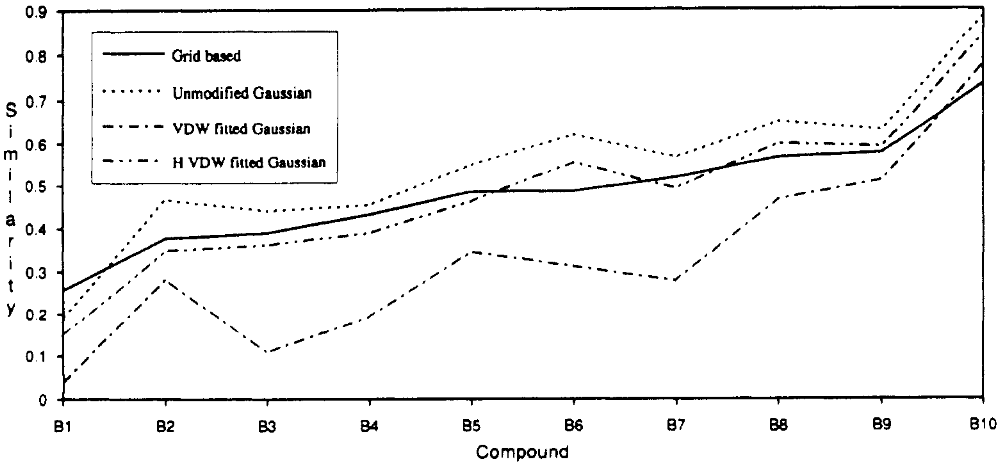


Figure 4. Similarity results for study 2.

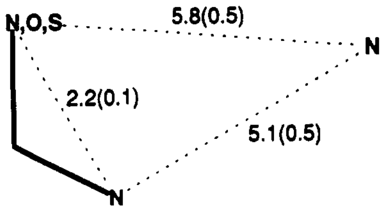


Figure 5. 3-D database search query used to obtain structures in study 3.

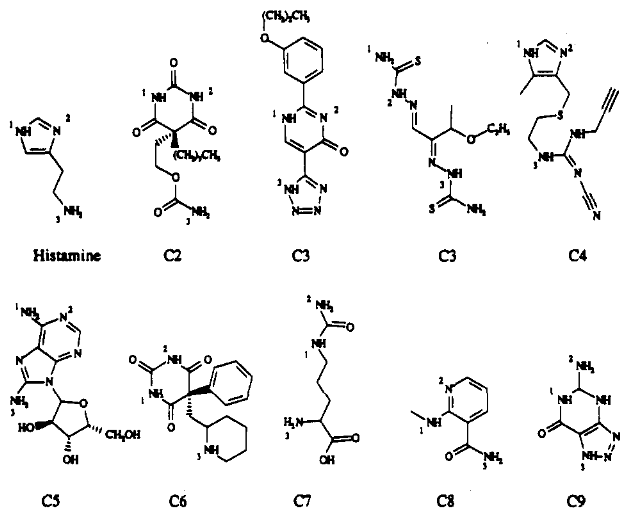


Figure 6. Structures used in study 3.

determined using all exponents.

Study 4. To compare the relative speeds of the similarity evaluation techniques, a large similarity calculation was devised. Two Np-apomorphine molecules (atomic coordinates available on request) as shown in Figure 8 were superimposed on each other. One of the molecules was then rotated 36 times in 10° increments about the *z* axis through the Np-apomorphine nitrogen atom. The similarity between the two molecules was determined after each rotation. The resulting similarity values obtained mirrored the trends shown in studies 2 and 3. One additional calculation was also undertaken with this investigation. Using the unmodified Gaussian functions, a distance cutoff of 3.6 Å was used, and all small exponent terms were removed. The similarity results for this evaluation system were always within 0.008 of those using all exponent terms with no distance cutoffs. The times required to carry out the similarity calculations are shown in Table III.

Study 5. One Np-apomorphine molecule (atomic coordinates available on request) as shown in Figure 8 was reoriented about each of the axes through its nitrogen atom. The resultant structure was then optimized against the Np-apomorphine molecule occupying its original position in space. In the first optimization, reorientation was achieved by applying 30° rotations and 1-Å translations in the *x*, *y*, and *z* axes through the Np-apomorphine nitrogen atom. In the second optimization, 20° rotations and 1-Å translations were applied in the *x*, *y*, and *z* axes through the Np-apomorphine nitrogen atom. The initial optimization calculations were executed using unmodified and H VDW fitted Gaussian functions, with and without a 3.6-Å distance cutoff. All calculations were run with the small exponent terms removed. For the first optimization, all calculations shifted the mobile molecule to the local minimum shown in Figure 9. When the rotation size was reduced to 20° for the second optimization system, all calculations shifted the mobile molecule back to it original position. Use of distance cutoffs approximately halved the time required for the optimization to finish (around 1 min required for optimization with distance cutoff on a PC 386-25 MHz).

DISCUSSION

Studies 1–3 show that, in general, the behavior of the Gaussian function similarity calculations closely mirrors that of the grid-based calculations. There are differences, however, when structural variation is limited. This is shown in Figure 2, where structures containing nitro groups are always more similar for Gaussian functions than for grid-based evaluations relative to the other analogues. The relative behavior of the various Gaussian functions is also different for this study. Figure 2 shows the VDW fitted functions most closely approximating the absolute similarity values produced by the grid-based evaluation. Figures 4 and 7 show the absolute H VDW fitted function similarity values closest to those of the grid-based system. In general, when compared with the earlier evaluation technique, the unmodified Gaussian functions produce higher absolute but similar relative results. The use of H VDW fitted functions results in a closer correspondence of absolute values. Utilizing all VDW fitted functions tends to produce a much higher sensitivity to dissimilarity than for the other systems (see Figures 4 and 7).

Removal of small exponent terms from the overlap evaluation functions has a negligible effect on the similarity results obtained (differences generally less than 1%). The same can be said for the use of distance cutoffs when the cutoff is set to greater than twice the van der Waals radius of the largest atom type in the system (see study 3).

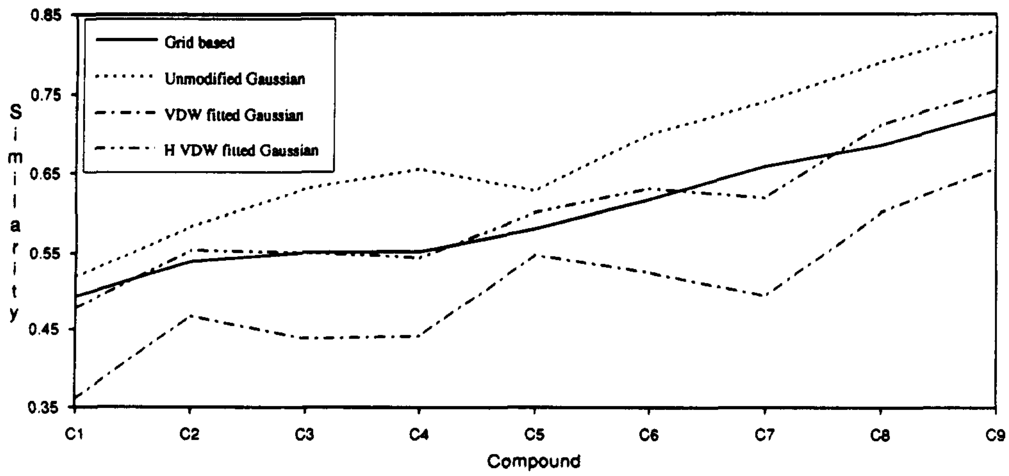


Figure 7. Similarity results for study 3.

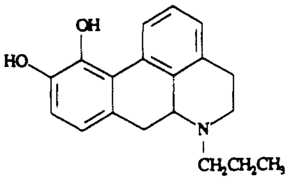


Figure 8. Structure used in studies 4 and 5.

Table III. CPU Time Taken for Study 4 Similarity Calculations

similarity calculation technique	time required for calcn (s)[a]
grid based	34 107
Gaussian	88
Gaussian with cutoff/less exponent terms	33

[a] Calculations executed on a VAX station 3520.

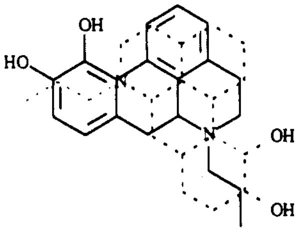


Figure 9. Local minimum found during first study 5 similarity optimization.

The CPU timings shown in Table III illustrate the benefits of the new Gaussian function calculations. Using Gaussian functions with no software tweaks, the similarity calculations are made almost 400 times faster than for the earlier grid-based evaluation method. With the removal of small exponent terms and the addition of a 3.6-Å distance cutoff, a thousandfold increase in speed is obtained with no significant loss in precision.

The general behavior of the unmodified and H VDW fitted functions with and without a 3.6-Å cutoff is found to be almost identical for optimization study 5. The use of a distance cutoff halved the time required for optimization to complete, again with no apparent impact on precision.

CONCLUSIONS

The use of Gaussian function approximations to atomic orbital electron density provides a rapid means for the evaluation of shape similarity.

It is not possible to say categorically which of the functions is the best for similarity evaluation. On the basis of best fit to the earlier grid-based evaluations, the H VDW fitted functions probably provide the best method. To maximize calculation speed, the removal of small overlap exponent terms and the addition of a distance cutoff are also recommended.

The results obtained with the new functions are generally found to be similar with those calculated by the earlier grid-based evaluation method. Differences do exist, however, especially for closely related systems with subtle functional groups differences. The use of the new technique for the evaluation of chiral coefficients as undertaken by Meyer,[9] therefore, requires further study to determine its suitability. Nevertheless, results obtained when studying widely varying molecular orientation and structure showed excellent correspondence with the earlier method. This augers well for the use of the new functions in both database searches and similarity optimizations.

ACKNOWLEDGMENT

This work is supported through an SERC CASE award with British Biotechnology Ltd., Oxford, England.

REFERENCES AND NOTES

(1) Carbo, R.; Leyda, L.; Arnau, M. An Electron Density Measure of the Similarity between Two Compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185–1189.
(2) Carbo, R.; Domingo, L. LCAO-MO Similarity Measures and Taxonomy. *Int. J. Quantum Chem.* **1987**, *32*, 517–545.
(3) 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.
(4) Bowen-Jenkins, P. E.; Richards, W. G. Quantitative Measures of Similarity between Pharmacologically Active Compounds. *Int. J. Quantum Chem.* **1986**, *30*, 763–768.
(5) Burt, C.; Richards, W. G. Molecular Similarity: The Introduction of Flexible Fitting. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 231–238.
(6) Burt, C.; Huxley, P.; Richards, W. G. The Application of Molecular Similarity Calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.
(7) Automated Similarity Package, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford on Thames, Oxford OX4 4GA, United Kingdom.
(8) Richard, A. M. Quantitative Comparison of Molecular Electrostatic Potentials for Structure-Activity Studies. *J. Comput. Chem.* **1991**, *12* (8), 959–969.
(9) Meyer, A. M.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 426–439.
(10) Manaut, M.; Sanz, F.; Jose, J.; Milesi, M. Automatic Search for Maximum Similarity between Molecular Electrostatic Potential Distributions. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 371–380.
(11) Mezey, P. G. Group Theory of Electrostatic Potentials: A Tool for Quantum Chemical Drug Design. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1986**, *12*, 113–122.
(12) Mezey, P. G. The Shape of Molecular Charge Distributions: Group Theory without Symmetry. *J. Comput. Chem.* **1987**, *8*, 462–469.
(13) Arteca, G. A.; Jammal, V. B.; Mezey, P. G. Shape Group Studies of Molecular Similarity and Regioselectivity in Chemical Reactions. *J. Comput. Chem.* **1988**, *9*, 608–619.
(14) Walker, P. D.; Arteca, G. A.; Mezey, P. G. A Complete Shape Group Characterization for Molecular Charge Densities Represented by Gaussian-Type Functions. *J. Comput. Chem.* **1990**, *12*, 220–230.
(15) Cioslowski, J.; Fleischmann, E. D. Assessing Molecular Similarity from Results of ab Initio Electronic Structure Calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64–67.
(16) Graham, M. S. Merck Sharpe & Dohme, Sea Program, *Q.C.P.E. 567*.
(17) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188–191.
(18) Hodgkin, E. E.; Richards, W. G. A Semi-Empirical Method for Calculating Molecular Similarity. *J. Chem. Soc., Chem. Commun.* **1986**, 1342–1344.
(19) Frisch, M. J.; Gordon, M.; Schlegel, H. B.; Raghavachari, K.; Binkley, J. S.; Gonzalez, C.; Defrees, D. J.; Fox, D. J.; Whiteside, R. A.; Seeger, R.; Melius C. F.; Baker, J.; Martin, R.; Kahn, L. R.; Stewart, J. J. P.; Fluder, E. M.; Topiol, S.; Pople, J. A.; *Gaussian 88*; Gaussian. Inc.: Pittsburgh, PA, 1988.
(20) Gill, P. E.; Murray, W. Algorithms for the Solution of Non-linear Least Squares Problems. *J. Numer. Anal.* **1978**, *15*, 977–992.
(21) Szabo, A.; Ostland, N. S. *Modern Quantum Chemistry*; Macmillan: New York, 1982; pp 410–412.
(22) Nelder, J. A.; Mead, R. Simplex Method for Function Minimization *Comput. J.* **1965**, *7*, 308–313.
(23) Elks, J.; Ganellin, C. R. *Chapman and Hall Dictionary of Drugs, Chemical data, structure and bibliography*; Cambridge University Press: Cambridge, U.K., 1990.
(24) CHEM-X, Chemical Design Ltd., Unit 12, 7 West Way, Oxford OX2 0JB, United Kingdom.