Utilization of Gaussian Functions for the Rapid Evaluation of Molecular Similarity

A. C. GOOD,† E. E. HODGKIN,‡ and W. G. RICHARDS*,†

Physical Chemistry Laboratory, Oxford University, South Parks Road, Oxford OX1 3QZ, United Kingdom, and British Biotechnology Ltd., Watlington Road, Oxford OX4 SLY, United Kingdom

Received November 22, 1991

An analytic method which permits the comparison of molecular electrostatic potential is introduced. The new technique fits the inverse distance dependence of the potential to an expansion of Gaussian functions and applies these functions to the Carbo similarity index. The major advantages of this approach are the very rapid computational speeds (2 orders of magnitude faster than previous methods) and superior convergence properties achieved during optimization of intermolecular electrostatic potential overlaps.

INTRODUCTION

Molecular similarity calculations are becoming a well-established modeling technique, with a number of different techniques proposed and applied.[1-16] They are useful both as parameters in quantitative structure–activity relationships and in finding the optimum position for superimposed structures by maximizing the similarity as a function of position.

As originally introduced by Carbo,[1,2] comparisons were made in terms of electron density, with the similarity R_{AB} between two molecules A and B given by

$$R_{AB} = \frac{\int P_A P_B \, d\nu}{\left(\int P_A^2 \, d\nu\right)^{1/2}\left(\int P_B^2 \, d\nu\right)^{1/2}}$$

where P_A and P_B are the electron densities of the two molecules being compared. The numerator term measures electron density overlap, and the denominator terms normalize the similarity results obtained. This formula has the virtue of being firmly based in quantum mechanics. Electron density, however, is not a very discriminating property, and for the purposes of medicinal chemistry, molecular electrostatic potential has the advantages of improved discrimination and ease of calculation.

A number of computer programs have been created which calculate molecular similarity in terms of electrostatic potential.[7-10] In the most widely used program of this type,[7] the electrostatic potentials used to evaluate similarity are calculated from point charges (q_i) assigned to each atom (i) such that the electrostatic potential at a point r for a molecule of n atoms is given by

$$P_r = \sum_{i=1}^{n} q_i/(r - R_i)$$

where R_i is the nuclear coordinate position of atom i. The electrostatic potentials are determined at the intersections of a rectilinear grid constructed around the two molecules. In order to avoid singularities at the atomic nuclei (where $1/r$ tends to infinity), the electrostatic potential is normally only determined outside the van der Waals volumes of the molecules in the calculation. These electrostatic potential values are then used to evaluate the Carbo formula (or the alternative created by Hodgkin[3]) numerically.

To gain computation speed the grid is normally coarse, with the unfortunate consequence that the integral is not well evaluated. In particular the optimization of similarity through the adjustment of relative molecular positions is coarse and crude. It is, for example, very difficult for the grid-based calculation to superimpose a molecule on top of itself, since the program tends to converge prematurely at some discrete point.

One improvement that can be made is the use of radial rather than rectilinear grids as implemented by Richard.[8] This technique offers a more efficient method for the evaluation of electrostatic potentials. Similarity evaluations of increased accuracy can thus be achieved through the use of a smaller grid increment.

In the present study the grid-based evaluation technique is replaced by number of analytic Gaussian functions so as to attain a fast and accurate method for similarity integral calculation.

GAUSSIAN FUNCTION APPROXIMATION TO $1/R$ CURVE

It is possible to substitute the $1/r$ term in the above equation with a Gaussian function approximation:

$$P_r = \sum_{i=1}^{n} q_i(\gamma_1 e^{-\alpha_1(r-R_i)^2} + \gamma_2 e^{-\alpha_2(r-R_i)^2} + \ldots \gamma_k e^{-\alpha_k(r-R_i)^2})$$

If this potential function is inserted into the Carbo formula we get the following equation

$$R_{AB} = \left[\sum_{i=1}^{n}\sum_{j=1}^{m} q_i q_j \int (G_1^i + G_2^i + \ldots G_k^i) \times (G_1^j + G_2^j + \ldots G_k^i) \, d\nu\right] \Big/ \left[\left(\sum_{i=1}^{n}\sum_{i=1}^{n} q_i q_i \left[\int (G_1^i + G_2^i + \ldots G_k^i)^2 \, d\nu\right]^{1/2}\right)\left(\sum_{j=1}^{m}\sum_{j=1}^{m} q_j q_j \left[\int (G_1^j + G_2^j + \ldots G_k^j)^2 \, d\nu\right]^{1/2}\right)\right]$$

where

$$G_k^i = \gamma_k e^{-\alpha_k(r-R_i)^2}$$

The integral terms expand into a series of two-center Gaussian overlap integrals. Consider a function containing two Gaussian terms

$$\int (\gamma_1 e^{-\alpha_1(r-R_i)^2} + \gamma_2 e^{-\alpha_2(r-R_i)^2})(\gamma_1 e^{-\alpha_1(r-R_j)^2} + \gamma_2 e^{-\alpha_2(r-R_j)^2}) \, d\nu = \gamma_1^2 \int e^{-\alpha_1(r-R_i)^2} e^{-\alpha_1(r-R_j)^2} \, d\nu + 2\gamma_1\gamma_2 \int e^{-\alpha_1(r-R_i)^2} e^{-\alpha_2(r-R_j)^2} \, d\nu + \gamma_2^2 \int e^{-\alpha_2(r-R_i)^2} e^{-\alpha_2(r-R_j)^2} \, d\nu$$

These two-center integrals have a simple form based on exponent values and distances between atom centers,[17] for example

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

It is therefore possible to break down the similarity computation in to a series of readily calculable exponent terms.

These terms can be evaluated extremely quickly, and since no singularity exists as r approaches zero, electrostatic potential

† Oxford University.
‡ British Biotechnology Ltd.

0095-2338/92/1632-0188$03.00/0

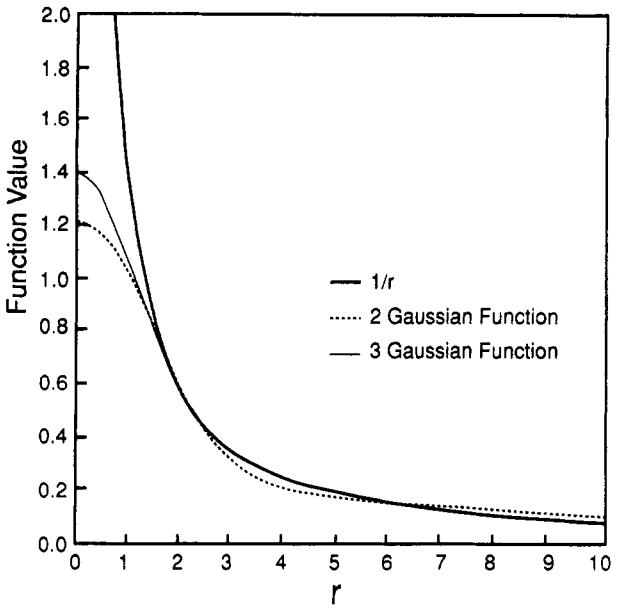

Figure 1. Gaussian function approximation to the $1/r$ curve.

evaluations need not be restricted to regions outside the van der Waals radii.

The number of Gaussian terms used in the inverse distance function depends on the accuracy of approximation that is required. Using the least-squares fit method,[18] the exponent and proportionality constants of two- and three-term Gaussian functions were optimized to fit the $1/r$ curve

$$\text{two Gaussians} - 0.2181e^{-0.0058r^2} + 1.0315e^{-0.2889r^2}$$

$$\text{three Gaussians} - 0.3001e^{-0.0499r^2} + 0.9716e^{-0.5026r^2} + 0.1268e^{-0.0026r^2}$$

The fit of these Gaussian functions to the $1/r$ curve is shown in Figure 1.

These functions were substituted into the Carbo formula and evaluated as shown above. The resultant exponent terms were placed in a subroutine which replaced the grid-based evaluation subroutines of the ASP program.[7] This subroutine executes a single point similarity calculation for a given set of coordinates and atomic point charges 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.[19]

SIMILARITY CALCULATIONS

To determine the behavior of the Gaussian functions two separate studies were undertaken.

In the first investigation a hypoglycemic active ring fragment[20] was compared with a number of related ring analogues.

The second investigation involved the similarity index optimization of an Np–apomorphine molecule against a second Np–apomorphine molecule occupying a different region of space.

For both investigations, the results obtained were compared with those produced by the grid-based electrostatic potential calculations of the ASP program.[7]

HYPOGLYCEMIC AGENT STUDY

Similarity calculations were applied to the hypoglycemic active lead fragment and its ring analogues shown in Figure 2. Molecules were built in Chem-X[21] using standard bond lengths and angles. The resulting structures were optimized in MOPAC5[22] using the PM3 parameter set, with the atomic point charges back-calculated to fit the molecular electrostatic potential.[23] The molecules were superimposed by a least-squares fit of atoms using the orientations shown in Figure 2. Single-point similarity calculations were then executed between the lead fragment and its analogues. The results were used to compare the similarity values produced by the Gaussian and grid-based functions.

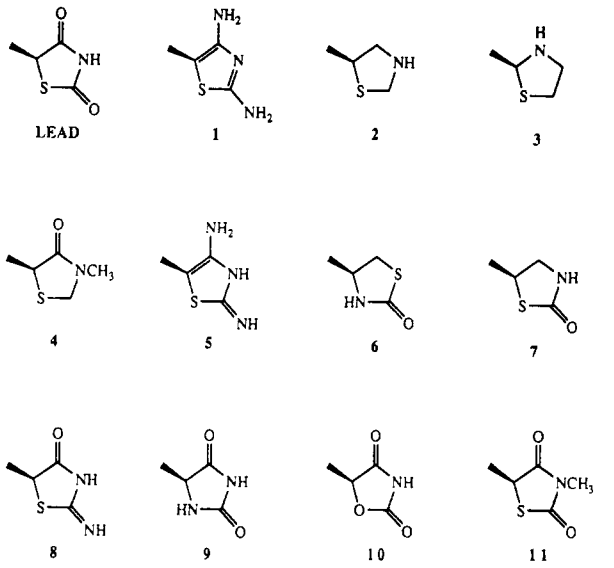

Figure 2. Hypoglycemic agent ring fragment and related analogues.

Table I. Similarity Results for Hypoglycemic Agent and Analogues

hypoglycemic agent fragment analogues	three-Gaussian function similarities	two-Gaussian function similarities	grid-based function similarities[a]
1	−0.089	−0.279	−0.415
2	−0.104	−0.189	0.000
3	−0.103	−0.124	0.049
4	0.413	0.431	0.196
5	0.267	0.246	0.230
6	0.383	0.332	0.491
7	0.536	0.541	0.579
8	0.804	0.821	0.699
9	0.898	0.904	0.853
10	0.968	0.973	0.933
11	0.982	0.986	0.939

[a] Default grid parameters used. Increment = 1.0 Å; extent = 4.0 Å.

Table I lists the results determined from these calculations (note that the fragments have been positioned in order of increasing grid-based function similarity). The best fit lines of each similarity evaluation method are shown in Figure 3.

NP–APOMORPHINE OPTIMIZATION STUDY

Similarity calculations were applied to the Np–apomorphine molecule shown in Figure 4. Structure building, optimization, and point charge calculation were determined as for the hypoglycemic agent study. The resultant structure was then shifted 1 Å and rotated 30° about the x, y, and z axes through the amine nitrogen atom. Finally the reoriented structure was optimized against the Np–apomorphine molecule in its original position in space. The results were used to compare the speed and convergence properties of the Gaussian and grid-based functions.

Table II shows the convergence and iteration times for the shifts and rotations applied to the moving Np–apomorphine during optimization, together with the final similarity results.

DISCUSSION

It is clear from the hypoglycemic ring analogue study that the general behavior of the Gaussian functions closely matches that of the grid-based similarity evaluation for fixed orientation

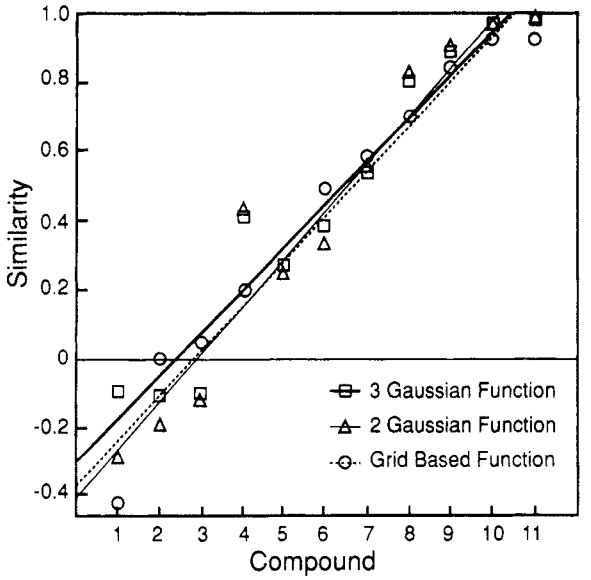

Figure 3. Similarity trends of the various methods of evaluation for the hypoglycemic agent ring fragment versus its ring analogues.

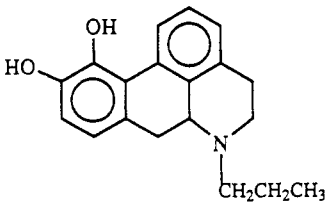

Figure 4. Np–apomorphine.

Table II. Similarity Results for Np–Apomorphine Simplex Optimization

simplex optimization results	three-Gaussian function calculations	two-Gaussian function calculations	grid-based function calculations[a]
simplex convergence CPU time, S[b]	253	114	9330
no. of simplex iterations	252	167	65
time per iteration,[b] s	1.0	0.7	143.5
final *x* rotation, deg	−30	−30	−19
final *y* rotation, deg	−30	−30	−26
final *z* rotation, deg	−30	−30	−24
final *x* translation, Å	−1.00	−1.00	−0.86
final *y* translation, Å	−1.00	−1.00	−0.79
final *z* translation, Å	−1.00	−1.00	−0.94
optimized similarity value	1.000	1.000	0.724

[a] Default grid parameters used. Increment = 1.0 A; extent = 4.0 Å. [b] Computation carried out on a Vaxstation 3520.

calculations. Figure 3 is a good illustration of this, with virtually identical similarity trends being exhibited for both similarity evaluation techniques.

Similar investigations have been carried out using H2 antagonist and prostaglandin ring fragments, with a correspondingly good agreement being found between the similarity results for the two methods.

The results for the similarity optimization of Np–apomorphine show the major advantages of the Gaussian functions over grid-based similarity evaluations. Both the two and three Gaussian functions manage to overlay the free moving Np–apomorphine structure back on top of the fixed Np–apomorphine molecule. The grid-based evaluation, however, converges before full molecular overlap has been achieved.

Similar studies were undertaken with the first five hypoglycemic agent ring analogues. One-Å shifts and 30° rotations were applied about the *x*, *y*, and *z* axes positioned at the ring centroid of each molecule. In all cases both the two- and three-Gaussian functions managed to achieve excellent molecular overlap, with the final similarity values always greater than 0.99. Every grid-based calculation converged prematurely into a local minimum, with all final similarity results less than 0.85.

The time taken to achieve convergence for the Gaussian functions is far less than that taken for the grid-based evaluation, with a 150–200-fold speed increase being achieved.

There appears to be little disadvantage in using the more approximate two-Gaussian function, since it produces results which correspond closely with those obtained with the three-Gaussian function. The speed of the two-Gaussian function is, however, significantly greater. This is because only three exponential terms need be evaluated during similarity value determination for the two-Gaussian function, as opposed to six for the three-Gaussian function.

CONCLUSIONS

These results show that speed increases of up to 2 orders of magnitude are achieved when Gaussian functions are used in place of grid-based evaluations in electrostatic potential similarity calculations. This speedup is achieved with no apparent loss in accuracy and a definite improvement in function sensitivity during optimization.

As with most optimizable properties, molecular similarity surfaces are made up of multiple minima. It is therefore quite possible that for systems where no obvious initial superposition exists, similarity optimizations will only yield a local optimum value. The new methodology presented here will not in itself overcome this problem. However, the rapidity with which the new functions' evaluations are calculated provides a possible answer, since it should now be possible to carry out Monte Carlo simulations or conformational analyses utilizing electrostatic potential similarity as the search variable. The use of Gaussian functions should therefore greatly enhance the flexibility of the calculations which may be undertaken with respect to the optimization of electrostatic potential overlap.

It is our intention to incorporate these modifications into the ASP program.[7]

ACKNOWLEDGMENT

This work is supported through a 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) 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.
(10) 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.
(11) Mezey, P. G. The Shape of Molecular Charge Distributions: Group Theory without Symmetry. *J. Comput. Chem.* **1987**, *8*, 462–469.

(12) 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.
(13) 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.
(14) Cioslowski, J.; Fleischmann, E. D. Assessing Molecular Similarity from Results of ab Initio Electronic Structure Calculations. *J. Am. Chem. Soc.* **1991**, *113*, 64-67.
(15) Graham, M. S. Merck Sharpe & Dohme, Sea Program, *Q.C.P.E. 567*.
(16) Meyer, A. M.; Richards, W. G. Similarity of Molecular Shape. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 426-439.
(17) Szabo, A.; Ostland, N. S. *Modern Quantum Chemistry*; Macmillan: Basingstoke, 1982; pp 410-412.
(18) Gill, P. E.; Murray, W. Algorithms for the Solution of Non-linear Least Squares Problems. *J. Numer. Anal.* **1978**, *15*, 977-992.
(19) Nelder, J. A.; Mead, R. Simplex Method for Function Minimization *Comput. J.* **1965**, *7*, 308-313.
(20) Takashi, S.; Katsutoshi, M.; Hiroyuki, T.; Yasuo, S.; Takeshi, F.; Yutaka, K. Study of Antidiabetic Agents. Synthesis of AL321 and Related Compounds. *Chem. Pharm. Bull.* **1982**, *30*, 3563-3573.
(21) Chem-X, Chemical Design Ltd., Unit 12, 7 West Way, Oxford OX2 0JB, United Kingdom.
(22) Stewart, J. J. P. MOPAC 5, *Q.P.C.E. 455*.
(23) Ferenzcy, G.; Reynolds, C. A.; Richards, W. G. Semi-Empirical AM1 Electrostatic Potential and AM1 Electrostatic Potential Derived Charges, a Comparison with ab Initio Values. *J. Comput. Chem.* **1990**, *11*, 159-159.

Compare_Conformer: A Program for the Rapid Comparison of Molecular Conformers Based on Interatomic Distances and Torsion Angles

ISTVÁN KOLOSSVÁRY† and WAYNE C. GUIDA*

Research Department, Pharmaceuticals Division, CIBA-GEIGY Corporation, Summit, New Jersey 07901

Received September 11, 1991

A computer program for comparison of the conformations of a number of related molecular structures is described. The comparisons are performed on either interatomic distances or torsion angles. The comparisons are accomplished on ordered pairs of distances or torsion angles, and the distance comparisons can be performed in a manner that allows permutation of the distance pairs being compared. The algorithm utilizes bit-string Boolean operations that allow the comparisons to be performed rapidly. The program should be useful for computer-assisted molecular modeling studies in which the viable conformers of bioactive analogues are compared in order to locate those conformers that place key substituents in the same spatial orientation.

INTRODUCTION

A central theme of computer-assisted molecular modeling has been the structural comparison of compounds that bind to the same molecular receptor. Such comparisons, particularly those among related structural analogues, have been employed to derive templates for drug design and to assist in the elucidation of receptor-ligand associations. A number of procedures have emerged that are of utility in molecular modeling studies. For example, a method termed the "Active Analog Approach" (AAA) pioneered by Marshall and coworkers[1] relies on the concept of a pharmacophoric pattern. This so-called pharmacophore is defined as the three-dimensional arrangement of atoms and/or functional groups essential for favorable receptor-ligand interactions. For compounds that lack structural rigidity, the potential pharmacophoric pattern for a particular molecule is not fixed but can vary substantially depending upon the conformational properties of the molecule under consideration. Thus, identification of the pharmacophore for flexible molecules is often an arduous task. The goal of a molecular modeling study involving AAA is to locate the pharmacophoric pattern common to conformers in a series of active molecules, and this goal can be achieved by coupling a systematic grid search of conformational space (without energy minimization) to orientation map calculations (identification of common intergroup distances).[2,3] An advantage of this method is that distance constraints derived from relatively rigid analogues can be used to limit the conformational search for more flexible analogues. The utility of AAA has been convincingly demonstrated with the suggestion of a unique bioactive conformation for a series of angiotensin-converting enzyme inhibitors.[4]

Numerous other methods have been employed for the identification of a common pharmacophore within a series of bioactive analogues[5-11] and for the location of molecules within three-dimensional databases that possess a particular pharmacophoric pattern.[12-16] All of these methods rely upon definition of the pharmacophore as a set of interatomic or intergroup distances and can be categorized (perhaps superficially) as distance geometry and/or pattern recognition approaches.

Of course, the identification of conformers sharing the same pharmacophoric pattern may be insufficient for complete rationalization of three-dimensional structure versus activity. Subsequent analysis may reveal whether steric and electrostatic properties of these conformers are optimal for binding. Volume mapping techniques can be used to provide additional steric information.[17,18] Coulombic and hydrophobic matching, which can be formulated in terms of electrostatic potentials and fields (potential gradient), may provide additional useful information concerning whether a particular conformer is likely to be active.[19-22] Furthermore, the approaches mentioned above rely on simplifying assumptions such as a single-binding mode. Nevertheless, these techniques have proven to be useful in the rationalization of three-dimensional structure-activity relationships.

Approaches such as AAA require collection of multiple conformational states of the analogues under consideration and comparison of interatomic or intergroup distances among the allowed conformations. In this paper, we describe an algorithm for the rapid comparison of a series of structural analogues, each of which may be associated with multiple conformers that may (or may not) have been previously subjected to energy minimization. We have developed a computer program that we call Compare_Conformer and abbreviate CP (for Com-

* Author to whom correspondence should be addressed.
† On leave from the Department of General and Analytical Chemistry, Technical University of Budapest, Szt. Gellért tér 4, H-1111 Budapest, Hungary.

0095-2338/92/1632-0191$03.00/0