Journal of Computer-Aided Molecular Design, 4 (1990) 231–238
ESCOM

J-CAMD 092

Molecular similarity: The introduction of flexible fitting

Catherine Burt and W. Graham Richards*

Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ, U.K.

Received 16 January 1990
Accepted 19 March 1990

Key words: Similarity; Conformational flexibility; Electrostatic potential; Electric field

SUMMARY

A method of comparing molecules on a quantitative basis which includes the possibility of compounds using their flexibility to achieve matching is described and applied to a series of hypoglycemic and hypolipidemic agents.

INTRODUCTION

Quantitative measures of molecular similarity are exciting interest in the pharmaceutical and agrochemical research fields. But existing programs compare given structures and do not allow for the obvious reality that molecules may be flexible. What is needed is an extension to current methodologies which permits a molecule to adapt its shape so as to mimic a lead compound, with the possibility of incurring a penalty function in doing so. Here we describe a program which has this capability and include an example of its application.

Carbo et al. [1] introduced a similarity index:

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

where ρ_A is the electron density of molecule A.

The numerator is a measure of the overlap of charge density for two superimposed molecules. The denominator is the normalizing factor. This formula is by no means unique and variations on

*To whom correspondence should be addressed.

the method of normalization can be adopted. One of these alternatives, which has been found to be a more satisfactory measure of molecular similarity was proposed by Hodgkin and Richards [2]:

$$H_{AB} = \frac{2 \int \rho_A \rho_B \, dv}{\int \rho_A^2 \, dv + \int \rho_B^2 \, dv}$$

The molecular similarity program ASP (Automatic Similarity Package) [3, 4] calculates molecular similarity indices based upon electrostatic potentials and electric fields, these parameters replacing the electron densities in the above formalisms.

The molecular similarity indices are very sensitive to the relative orientations of the two compounds. This gives rise to the need for a method of optimization of the molecular similarity indices. Often the nature of a macromolecular receptor is unknown and must be inferred from the properties of the molecules that bind to it. Molecular graphics packages can be used to superimpose active molecules on a screen. There remains the difficulty of deciding just how to superimpose two molecules which may have very little in common in terms of obvious binding interactions. Quantitative molecular similarity can provide the answer by adjusting the relative positions and conformations to maximize similar features.

THEORY

The molecular similarity index is optimized using the SIMPLEX method of Nelder and Mead [5]. The simplex method minimizes a function of n variables by comparison of the function value at the n + 1 vertices of a general simplex, followed by replacement of the highest value by another point. Hence in order to maximize the similarity index, it is necessary to minimize the negative values. The highest and lowest values of the negative molecular similarity indices are determined. The centroid of all the points but the highest is determined and the highest value is 'reflected' through this 'centre of gravity' and the molecular similarity index is calculated at this point and compared with the other function values. Depending on the outcome of this test, the new point is accepted or rejected and a further expansion or contraction made. When no further progress can be made the sides of the simplex are reduced in length and the process is repeated. Since the molecular electrostatic potential and molecular electric field molecular similarity indices are computed rapidly, the simplex method, which is also robust, has been found to be an effective method for the maximization of the molecular similarity index.

Molecular similarity indices can be optimized within ASP by fixing the lead molecule and translating and rotating in space the molecule whose similarity index is to be optimized until it lies in its position of maximum similarity with respect to the lead molecule. In this case there are 6 variables in the simplex, the x-, y- and z-translations and x-, y- and z-rotations. The relative conformations as well as positions of a given series of compounds are important when comparing their similarity indices. The novel feature is that the similarity indices may also be optimized by including rotations about the torsional bonds of the 'moving' molecule. In this case the number of variables

in the simplex is 6 plus the number of torsional rotations. The torsions can be perceived by the program or defined by the user. The criteria for a rotatable bond in the perception routines is that neither atom of the bond should be a terminal atom nor a member of the same ring system. If the user wishes to restrict rotations to a limited number of torsions or for systems containing non-terminal double bonds, the torsions should be user defined. The first atom of the first torsion may be located anywhere within the structure, but further torsions should be defined eminating along each branch of the structure in the order they occur and in the same direction. The first vertex of the simplex corresponds to the molecule in its initial orientation before any translations or rotations are performed. The other vertices of the simplex are generated by applying random translations and rotations to the molecule as a whole and random rotations about each torsion to give a new orientation of the molecule corresponding to another vertex of the simplex. The translations and rotations of the molecule as a whole compensate for any arbitrariness in the definition of the first atom of the first torsion. Each random translation and rotation is applied to the 'moving' molecule a total of n times so that an n + 1 simplex is produced.

It is possible, however, to produce conformations which, while giving a very large degree of similarity, are energetically unrealizable. Hence, the molecular similarity index may be weighted to a Boltzmann factor so that a conformation that the molecule could never adopt is not produced as an optimum.

$$\text{weighted index} = \text{actual index} \times e^{-(c\,\Delta E/RT)}$$

where c = weighting factor, ΔE = energy of rotated conformation – energy of initial conformation, R = gas constant, and T = temperature.

The weighted index is optimized rather than the actual similarity index. If the energy difference between the rotated and initial conformations is negative then the above formalism could produce a molecular similarity index that is less than the original one as maximum. Since the aim is to optimize the similarity index, ΔE is set to zero if it is found to be negative. ΔE comprises a coulombic term and a van der Waals term and the parameters are taken from Allinger and Yuh's MM2 program [6]. In this way an energy difference between conformational states can be rapidly computed but should be regarded as a guideline as to whether a particular conformation could be adopted rather than an accurate measure of the energy difference between the conformations.

The weighting factor is necessary so that the actual value of the similarity index is weighted to the energy difference in order to reduce its value by an appropriate amount. To illustrate let us consider the reduction in value of a similarity index of 0.7 for various energy differences and weighting factors (see Table 1). When the value of the weighting factor c is 1, the molecular similarity index falls off rapidly. In contrast, when c = 0.01 and the energy difference between the conformations is 10 kcal/mol, a molecular similarity index of 0.7 is only reduced to 0.591. The Boltzmann distribution shows that for an energy difference of 3 kcal/mol only a few molecules in a thousand will achieve the conformation. However, this does not take into account any favorable interactions due to the binding of the ligand to the receptor. An intermediate value of the weighting factor, c = 0.1, was therefore chosen as the default value for ASP.

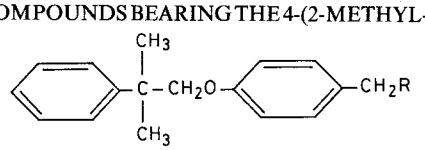
APPLICATION

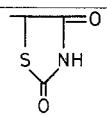
Takashi et al. [7] have synthesized a series of compounds bearing the 4-(2-methyl-2-phenylpropoxy)benzyl moiety and evaluated their hypoglycemic and hypolipidemic activities with genetically

TABLE 1
THE WEIGHTED VALUES OF A SIMILARITY INDEX OF 0.7 FOR VARIOUS ENERGY DIFFERENCES AND WEIGHTING FACTORS c

ΔE (kcal/mol)	Weighted index values				
	c = 1	c = 0.15	c = 0.1	c = 0.05	c = 0.01
1	0.129	0.534	0.591	0.643	0.688
2	0.024	0.422	0.499	0.591	0.677
3	4.4×10^{-3}	0.328	0.422	0.543	0.665
4	8.2×10^{-4}	0.254	0.356	0.499	0.654
5	1.5×10^{-4}	0.197	0.301	0.459	0.643
6	2.8×10^{-5}	0.153	0.254	0.421	0.633
7	5.2×10^{-6}	0.119	0.215	0.388	0.622
8	9.6×10^{-7}	0.092	0.181	0.356	0.612
9	1.8×10^{-7}	0.071	0.153	0.328	0.601
10	3.3×10^{-8}	0.056	0.129	0.301	0.591

TABLE 2
BIOLOGICAL PROPERTIES OF COMPOUNDS BEARING THE 4-(2-METHYL-2-PHENYLPROPOXY) BENZYL MOIETY



Compound label	R	Activity	
		Hypoglycemic activity	Plasma triglyceride lowering activity
Lead		3	3
1	—CHCOOEt Cl	2	4
2	—CHCOOEt SH	2	0
3	—CHCOOEt $SCOCH_3$	1	0
4	—CHCOOH SCH_3	1	0
5	—CHCOOH $SCH_2CH(NH_2)COOH$	1	0
6	—CHCOOEt $SCH_2CH_2NH_2$	1	0

obese and diabetic mice, yellow KK. Some of the compounds together with their activities are shown in Table 2.

The compounds were built within the molecular modelling package CHEMX [8]. The 4-(2-methyl-2-phenylpropoxy)benzyl moiety is common to all structures and was replaced by a CH_3 group so as to simplify the calculations. The atomic charges were calculated and the geometries were optimized using the AMPAC [9] program together with the AM1 [10] hamiltonian. The straight chain analogues were superimposed on the lead compound by a least squares fitting of atoms. Carbo electrostatic potential molecular similarity indices were optimized with respect to the position of the lead compound using the program ASP with 3 different sets of criteria.

Optimization 1: The similarity indices were optimized by translation and rotation of the molecules in space with no rotation about torsional bonds.

Optimization 2: The similarity indices were optimized by translation and rotation of the molecules in space and full rotation about all torsional bonds. The molecular similarity index is not weighted to a Boltzmann factor during the optimization.

Optimization 3: The similarity indices were optimized by translation and rotation of the molecules in space and full rotation about all torsional bonds. The molecular similarity index is weighted to a Boltzmann factor with a weighting factor of 0.1 during the optimization.

TABLE 3
ORIGINAL AND OPTIMIZED VALUES OF THE CARBO ELECTROSTATIC POTENTIAL MOLECULAR SIMILARITY INDEX WITH RESPECT TO THE LEAD COMPOUND

Compound label	Original index	Optimized value of the index					
		Optimization 1		Optimization 2		Optimization 3	
1	0.264	0.685[a]		0.788[a]		0.641[a]	
		(72)[b]	(159)[b]	37.825[c]	(151)[b]	−3.000[c]	
2	0.338	0.659		0.705		0.656	
		(46)	(125)	6.293	(138)	−0.276	
3	0.611	0.681		0.825		0.785	
		(49)	(138)	−1.865	(168)	−0.950	
4	0.651	0.692		0.727		0.719	
		(44)	(92)	0.208	(114)	−1.245	
5	0.545	0.642		0.804		0.778	
		(53)	(192)	3.868	(247)	−0.055	
6	0.312	0.620		0.679		0.662	
		(57)	(135)	382.904	(335)	0.005	

Optimization 1: Optimization by translation and rotation in space with no rotation about torsional bonds.

Optimization 2: Optimization by translation and rotation in space and rotation about torsional bonds. The molecular similarity index is not weighted to a Boltzmann factor during the optimization.

Optimization 3: Optimizaton by translation and rotation in space and rotation about torsional bonds. The molecular similarity index is weighted to a Boltzmann factor with a weighting factor of 0.1 during the optimization.

[a]Optimized value of index.

[b]Number of iterations of the simplex loop.

[c]Energy difference between final and initial conformations (kcal/mol).

RESULTS AND DISCUSSION

Table 3 shows the original values of the Carbo electrostatic potential molecular similarity indices together with their optimized values for each method of optimization. Also shown are the number of iterations of the simplex loop before convergence was obtained and the energy differ-

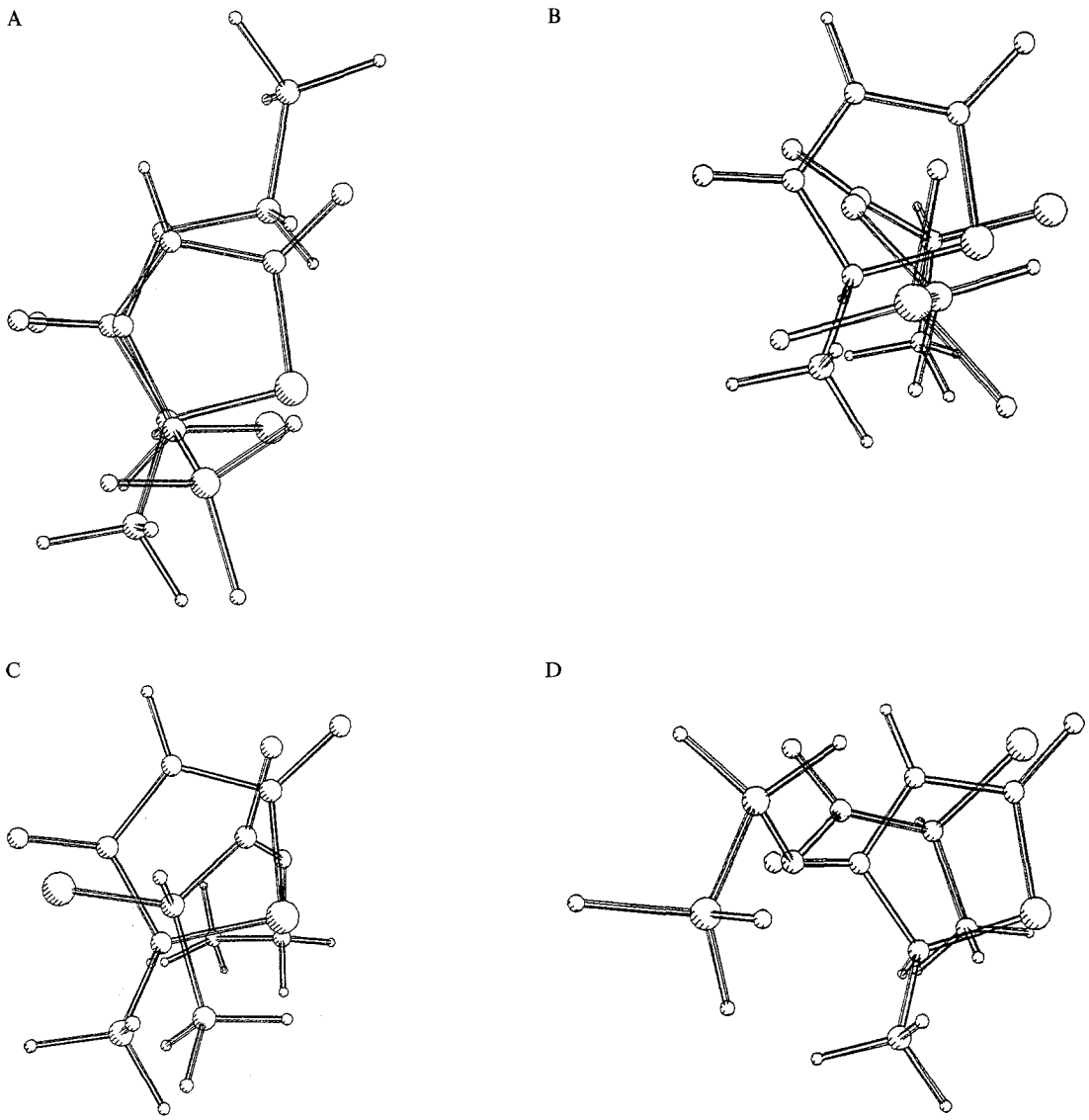


Fig. 1. (A) Original orientation of structure 1 with respect to the lead compound before optimization was performed; (B) Relative orientation following optimization by method 1; (C) Relative orientation following optimization by method 2; (D) Relative orientation following optimization by method 3.

ence between final and initial conformations if rotations about the torsional bonds have been performed. The optimized values given by optimization 2, as expected, have the largest values since there are more degrees of freedom in the simplex and there is no energy weighting to damp down the high energy dihedral rotations. When the weighting factor of 0.1 is introduced for optimization 3, again we would hope that the values obtained using this method of optimization would be greater than those obtained by optimization 1. This is true for 4 out of the 6 compounds, compounds 1 and 2 being exceptions although the difference in the values obtained by the different methods is only very small ($\Delta = 0.044$ and 0.003, respectively). For optimizations performed with no energy weighting 3 compounds (1, 2 and 6) are in clearly energetically unacceptable conformations (i.e. have an energy difference greater than 4 kcal/mol for the optimum position with respect to the initial conformation). This is due to a steric clash of atoms.

It is beyond the scope of this paper to list the change in conformation and the relative motion of each structure with respect to the lead compound for each of the methods of optimization, but structure 1 will be used as an illustration. Figure 1A shows the original orientation of structure 1 with respect to the lead compound before optimization was performed. Figure 1B shows the relative orientation following optimization by method 1. Structure 1 has been translated in the x, y and z directions by approximately 0.6, -0.7 and 0.02 Å and rotated about the x, y and z axes by -0.391, -0.263 and -1.656 radians respectively. Figure 1C shows the relative orientations of the two molecules following optimization using method 2. The relative translations and rotations are -0.3Å, 0.5Å, 0.1Å, 2.517 rad, 0.598 rad and -1.325 rad. In addition, the conformation of the molecule has altered considerably with rotations about torsional bonds varying in magnitude from approximately 6 to 149 degrees. However, the conformation is energetically unfavorable due to a steric clash of alkyl groups. The relative orientation of the two molecules following optimization by method 3 can be seen in Fig. 1D. Structure 1 has been translated by -0.2Å, -0.2Å and -1.1Å and rotated by -0.528, 0.131 and -1.480 radians. Again the conformation of the molecule has altered considerably with rotations about torsional bonds varying from 12 to 125 degrees. It can be seen that there are no obvious steric clashes in this conformation.

The correlation of biological activity with the molecular similarity indices is less clear-cut. Optimization by all methods raises the value of the similarity index for all structures above the value of 0.6 whereas an arbitrary superposition gives a wide spread in the values of the molecular similarity indices that can in no way be related to the biological activity. This illustrates the fact that the value of the molecular similarity index is very sensitive to the relative orientations of the two molecules and that optimization can play a valuable role in molecular superposition. One reason for the lack of correlation with biological activity could be that the 4-(2-methyl-2-phenylpropoxy) benzyl moiety has been replaced by the less bulky methyl group and that optimization has moved the molecule as a whole so that this group would occupy a position that sterically it could not adopt at the active site of the receptor. The AM1 calculations are not parameterized for the sulphur atom and depend upon Mulliken population analysis [11]. A better correlation might be obtained using point atomic charges derived from the PM3 [12] electrostatic potential. The reasons for a less than impressive correlation of the similarity index with biological activity in this particular case could stem from the choice of parameters or, more likely, from the nature of the biological data. Our intention has, however, been to introduce the methodology rather than present a striking demonstration.

CONCLUSION

A prescription for the optimization of molecular similarity indices incorporating flexible fitting would seem to be as follows:

(i) The optimization should be performed with no weighting of the similarity index to the energy difference between the conformations. This will give the largest possible value of the similarity index within the limitations of the SIMPLEX method.

(ii) If the above method of optimization produces a compound whose conformation is energetically unrealizable as maximum, the molecular similarity index should be re-optimized using different values of the weighting factor c, starting at low values and gradually increasing, until a conformation with a realizable energy difference is obtained.

Just how much energy a molecule can expend in rotation about torsional bonds will depend upon the binding energy at the active site. Since this information, in general, will be unavailable the weighting factor c should be a variable parameter left at the user's disposal. With the addition of conformational flexibility in matching structures the role of molecular similarity in structure–activity should be much more widely applicable.

ACKNOWLEDGEMENTS

The authors thank Drs. C.M. Edge and R.M. Hindley of Beecham Pharmaceuticals for helpful discussions.

REFERENCES

1 Carbo, R., Leyda, L. and Arnau, M., Int. J. Quantum Chem., 17 (1980) 1185.
2 Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem. Quantum Biol. Symp., 14 (1987) 105.
3 A.S.P. (Automatic Similarity Package), Oxford Molecular, Terrapin House, South Parks Road, Oxford OX1 3UB, U.K.
4 Burt, C., Huxley, P. and Richards, W.G., J. Comput. Chem., in press.
5 Nelder, J.A. and Mead, R., Comput. J., 7 (1965) 308
6 Allinger, N.L. and Yuh, Y.H., MM2, Q.C.P.E. 395, Indiana University Chemistry Department, Bloomington, IN, U.S.A.
7 Takashi, S., Katsutoshi, M., Hiroyuki, T., Yasuo, S., Takeshi, F. and Yutaka, K., Chem. Pharm. Bull., 30 (10) (1982) 3563.
8 CHEMX, Chemical Design Ltd., Unit 12, 7 Westway, Oxford OX2 0JB, U.K.
9 Dewar, M.J.S. and Stewart, J.J.P., Q.C.P.E. Bull., 6 (1986) 24, QCPE 506, AMPAC.
10 Dewar, M.J.S., Zoebisch, E.G., Healy, E.P. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
11 Mulliken, R.S., J. Chem. Phys., 23 (1955) 1833.
12 Stewart, J.J.P., J. Comput. Chem., 10 (1989) 209, 221.