CROATICA CHEMICA ACTA CCACAA **71** (3) 689–696 (1998)

ISSN-0011-1643
CCA-2520 *Original Scientific Paper*

A Comparative Study of Electrostatic Similarity Indices

Črtomir Podlipnik and Jože Koller*

Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1001 Ljubljana, Slovenia (e-mail: crtomir.podlipnik@uni-lj.si)

Received August 11, 1997; revised February 10, 1998; accepted February 12, 1998

In this study, the ethanol molecule was superimposed on a set of molecules (ethane, ethantiol, ethilamine, methanol, propane and water). Similarity, on the basis of the molecular electrostatic potential (MEP), was evaluated on the semiempirical AM1 and HF/6–31+G* . The Carbo index was used as a measure of similarity and three different formalisms were used for similarity evaluations: the MEP by the grid method, the ESP charge by the grid method and the ESP charge by the Gaussian approximation. It was found that the choice of the MO method and formalism do not play an important role when indices are high. When we superimpose two electrostatic potentials of molecules that are not similar, as for example electrostatic potentials of ethanol and ethane, the Gaussian approximation gives evidently higher values of the similarity indices than the grid method.

INTRODUCTION

Molecular similarity studies have become the focus of intense scientific interest in recent years.[1,2] Many practical applications of molecular similarity have been used in various fields of chemical science: drug design,[3–7] selection of analogs for chemicals,[8] and estimation of molecular properties.[9,10] In drug design, similarity/dissimilarity based methods have been very useful in rational selection of candidates from large databases.[11–14] The use of molecular similarity methods is based on the structure-property similarity principle.[15] This notion states that similar structures usually have similar

* Author to whom correspondence should be addressed.

properties. The main aim of the molecular similarity studies is design of new quantitative measures of molecular similarity. Many of these approaches use distance or angular information to define the degree of resemblance between two molecules. The similarity between a pair of molecules may then be calculated based on the overlap of the corresponding fields of the two molecules, using a similarity coefficient such as the Carbo index,[16] a form of the long-established cosine coefficient. There are many properties that we may compare in these similarity studies such as electrostatic potential,[17–20] electron density[16,21,22] and molecular lipophilicity potential.[23] In this study, we have focussed on the different approaches for electrostatic similarity evaluation between small molecules. Electrostatic potential is one of the most important molecular descriptors. Many studies show the importance of electrostatic interaction between organic and bioorganic molecules.[24–27] We have compared the results of similarity evaluation that were determined using two different approaches: the numeric grid method and the analytical approach which uses Gaussian approximation.[20]

METHODS

We used Spartan 4.0 for building molecular structures, their model energy minimization,[28] calculation of the molecular electrostatic potential (MEP), and the charges from the molecular electrostatic potential (ESP charges). Energy minimization and calculation of the MEP from the molecular wave function were calculated both semiempirically by the AM1 method[29] and HF/ 6–31+G*. The ESP charges were then calculated from MEP by least square fitting.[30]

We used cumulative indices for the calculation of the overall molecule similarity between two overlaid molecules. These indices are called cumulative because the overall similarity is determined through the accumulation of property overlap or difference values over the whole space. The most widely used form of cumulative index applied for the calculation of molecular similarity was proposed by Carbo *et al.*:[16]

$$R_{AB} \frac{\int P_A P_B d\tau}{(\int P_A^2 d\tau)^{1/2} \cdot (\int P_B^2 d\tau)^{1/2}}$$

The molecular similarity index R_{AB} is determined from the structural properties P_A and P_B of the two molecules being compared. In this work, the Carbo index was used for quantitative estimation of the molecular similarity based on electrostatic potential. The most common procedure for calculating the similarity index is the grid method. A molecule is positioned at the centre of a 3–D grid and then the molecular property is calculated at each point of the grid. The similarity between a pair of molecules is then es-

timated by comparing the corresponding properties at each grid point and summing up over the entire grid. Results are normalized with a suitable normalizing factor to put the resulting similarities into the range from –1.0 to +1.0, where –1.0 corresponds to total complementarity and +1.0 to total similarity between molecules. This numerical approach contains certain weaknesses, the biggest being the low speed of similarity index computation.

Good *et al.*[20,31] proposed an alternative approach in which the property distribution is approximated by the sum of the Gaussian functions that can be processed analytically. These analytical evaluations were found to be orders of magnitude faster (with only minimal effect on its accuracy) than the equivalent numerical calculations.

We used three different approaches for similarity indices evaluation:

(A) Numeric similarity calculation with the Spartan 4.0 module. This module enables grid method similarity calculations and uses simplex optimization[32] to find out the maximal similarity index. We used two different potentials for similarity index calculation: (**A1**) MEP, which is accessed directly from the wave function and (**A2**) potential derived from the ESP charges. Molecular similarity indices were calculated using AM1 and HF/6–31+G* geometry. The grid points inside the van der Waals envelope of the molecule were excluded from these calculations.

(B) Analytic similarity calculations that use Gaussian approximation (GA) with the SimMol[33] program. We used this approach to compare potentials from AM1 and HF/6–31+G* ESP charges of the set of molecules and ethanol molecule. The simplex optimization was used to determine the relative orientation of two molecules that corresponds to the highest similarity.

(C) Numeric similarity calculation within the SimMol program. Input for the program are the ESP charges, geometries and relative orientations of the pair of molecules. The relative orientations are the results of calculations using approach B. We have introduced factor ζ, by which the van der Waals atomic radii have been multiplied. In this approach, we used two different models for potential from ESP charges: **(C1)** We used the same model for electrostatic potential as in approach A2. (**C2**) We expanded the potential in terms of the linear combination of three Gaussians functions like in approach B. Both models and different values of factor ζ were used for single point similarity indices calculation. We used these two models to compare potentials that arise from AM1 ESP charges of molecules.

RESULTS AND DISCUSSION

Numerical (grid) evaluation of similarity indices from the MEP (A1 in Table I) is a time consuming process; however, the theory behind this calculation is very clear and less approximate than all the other calculations presented in this paper. This is the reason why we have used results of this

TABLE I

Superposition of an ethanol molecule on a set of molecules. The Carbo index is evaluated on the basis of the AM1 and HF/6–31+G* electrostatic potential. Last row: the CPU time required for computations. LEGEND: A1 – MEP were used for the similarity indices calculation by the grid method within SPARTAN similarity module; A2 – ESP charge potentials were used for calculation of similarity indices with grid method (SPARTAN); B – Analytical approach that uses Gaussian approximation of ESP charges potentials for calculation of the similarity index. These calculations were made with SimMol program.

	A1		A2		B	
	AM1	6–31+G*	AM1	6–31+G*	AM1	6–31+G*
Methanol	0.988	0.980	0.976	0.990	0.968	0.995
Water	0.981	0.961	0.984	0.973	0.943	0.955
Ethanthiol	0.975	0.932	0.994	0.967	0.989	0.972
Ethylamine	0.955	0.911	0.939	0.837	0.913	0.888
Propane	0.167	0.435	0.369	0.281	0.727	0.515
Ethane	0.201	0.310	0.337	0.235	0.736	0.397
CPU time / s	780	3767	99	126	2.9	2.9

type of calculation as a reference when comparing different approaches of similarity index evaluation. The time required for numerical calculation of the similarity index from MEP by the grid method is very long and it strongly depends on the basis set used for calculation of the electrostatic potential (last row in columns A1 in Table I). If we reduce the amount of initial information used for similarity index evaluation from the MEP to the ESP charges potential (column A2 in Table I), some differences between indices values are obtained. These differences tended to be larger when two dissimilar molecules were compared. The mentioned approach is approximately 10 times faster than A1 but it is practical only when potentials of relatively small molecules (small grids) are compared. In the next approach, we approximate the ESP charges potential with the sum of the three Gaussian functions (column B in Table I). Gaussian approximation enables a fast analytical calculation of the similarity index. These calculations are approximatelly two to three orders of magnitude faster than the corresponding calculations based on the grid method (last row in Table I). We found that for similar molecules the difference between indices obtained by A2 and B is small, whereas, when potentials of two molecules are different (low similarity indices), this differences could be very large. It is shown in Table I that when a pair of molecules has similar electrostatic potentials, then the value of the similarity index depends less on the level of the electrostatic potential calculation (AM1 or HF/6–31+G*) than when we compare two electrostatic dissimilar molecules (*e.g.* ethanol and ethane). From the charge distribution analysis of ethane it can be seen that charges on carbon atoms have a different sign at AM1 (–0.120) than at HF/6–31+G* (0.026). This could be the rea-

son for the differences in comparing electrostatic potentials of ethane and ethanol at various levels of electrostatic potential calculation.

In this study, we tried to find out the reasons for the large differences between the similarity indices obtained from A2 and B when we compared two dissimilar ESP charge potentials. For this reason, the indices that were obtained using approach B from AM1 ESP charge potential were recalculated by C1 and C2 methods using the SimMol program. Factor ζ which is introduced in C1 and C2 methods determines to what extent the van der Waals volume of the molecule is excluded from the grid method calculations. From Figure 1 it is evident that, in the case of ethanol and ethane which have different electrostatic potentials, the similarity indices calculated using both approaches C1 and C2 rapidly rise with the reduction of factor ζ (from $R_{AB} \approx 0.2$ for $\zeta = 1.0$ to $R_{AB} \approx 0.7$ for $\zeta = 0.2$). The differences between C1 and C2 values of the Carbo index at the same factor ζ are small. This means that the main reason for the differences between A2 and B values of the similarity index is probably not the quality of the fitting (three Gaussian) function. Comparing the results obtained by different approaches, we found that the results of approach C1 for $\zeta = 1$ are the same as in the numerical approach A2, when all the points inside the van der Waals spheres were excluded. The results of approach C2 are an approximation of the results of approach B when all the grid points are included, that is, in the case when $\zeta = 0$. We may conclude that the source of the difference between A2

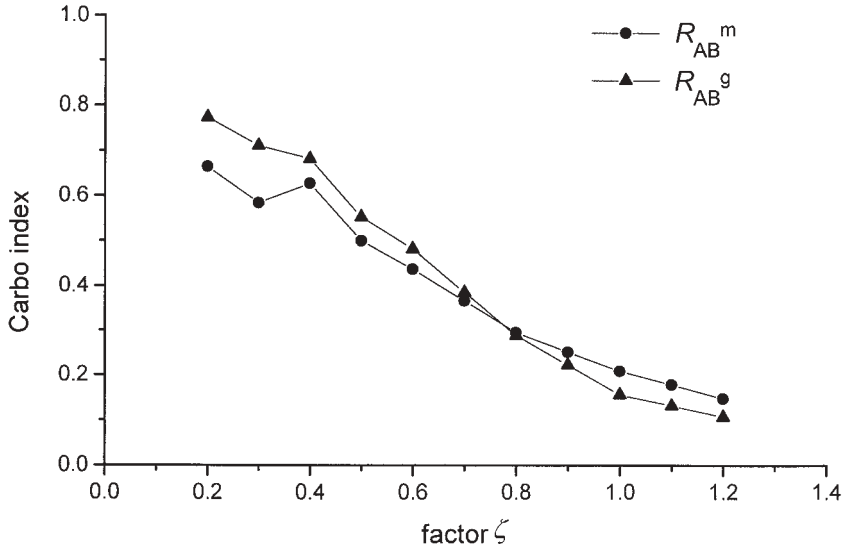

Figure 1. Influence of factor on the similarity indices for superposition of the electrostatic potentials of ethanol and ethane. Legend: $R_{AB}{}^{m}$ – the Carbo index calculated by C1 approach. $R_{AB}{}^{g}$ – the Carbo index calculated by C2 approach.

and B values of the similarity index is in the ways of accounting the potential inside the van der Waals spheres. The two other indices (Hodgkin[17] and Petke[34]) were also used for similarity measures. It can be seen that the found discrepancy did not depend on the chosen similarity measure.

CONCLUSIONS

Evaluation of similarity indices from the MEP is a time consuming process; however, the theory behind this calculation is very clear and less approximative. Grid calculations from ESP charges are much faster than similarity indices calculations from MEP, but they are still too slow for treating larger molecular systems. The Gaussian approximation enables fast analytical calculation of the similarity index; it speeds up evaluations of electrostatic similarity by two to three orders of magnitude. The Gaussian approximation and the grid method give comparable values of indices if two molecules have a similar shape of the electrostatic potential. If we compare two molecules with different potential shapes, a big difference between the indices calculated using the grid method and the Gaussian approximation is observed. Our results show that the main reason for the differences between indices calculated from ESP charge potential using the grid method and the Gaussian approximation is the way in which the potential inside the atomic van der Waals radii is taken into account. Since the numerical grid method excludes the grid points inside the atomic van der Waals spheres (problems with singularity), the Gaussian approximation is not restricted to regions outside the atomic van der Waals radii. If we take into consideration that ESP charge potential is well defined only at points outside the atomic van der Waals spheres of molecule which define the van der Waals surface, then the grid method results are more relevant than those determined by the Gaussian approximation. However, the Gaussian approximation offers a very fast and relatively accurate solution for the similarity index calculation. Good et al. also showed that, when used as parameters in QSAR calculations, analytically derived similarity inidices were more predictive than their numerical counterparts.[5,6]

Acknowledgements. – This work was supported by the Slovene Ministry of Science and Technology. Many thanks go to prof. Dušan Hadži for his help on our project and hours of fruitful discussions.

REFERENCES

1. G. M. Maggiora and M. A. Johnson (Eds.), *Concepts and Application of Molecular Similarity* , John Wiley and Sons, 1990.
2. P. M. Dean (Ed.), *Molecular Similarity in Drug Design*, Blackie Academic & Professional, 1995.

3. P. M. Dean, P. -L. Chau, and M. T. Barkat, *J. Mol. Struct.* **256** (1992) 75–89.
4. A. Seri-Levy and W. G. Richards, *Tetrahedron-Asymmetr.* **4** (1993) 1917–1921.
5. A. C. Good, S. So, and W. G. Richards, *J. Med. Chem.* **36** (1993) 433–438.
6. A. C. Good, S. J. Petterson, and W. G. Richards, *J. Med. Chem.* **36** (1993) 2929–2937.
7. G. Klebe, U. Abraham, and T. Mietzner, *J. Med. Chem.* **37** (1994) 4130–4146.
8. S. C. Basak and G. D. Grunwald, *SAR and QSAR in Environ. Res.* **2** (1994) 289–307.
9. S. C. Basak, B. D. Gute, and G. D. Grunwald, *Croat. Chem. Acta.* **69** (1996) 1159––1173.
10. R. Carbo, E. Besalu, Ll. Amat, and X. Fradera, *J. Math. Chem.* **18** (1995) 237–246.
11. Bath, A. Poirette, P. Willet, and F. H. Allen, *J. Chem. Inf. Comput. Sci.* **34** (1994) 141–147.
12. V. J. van Geerestein, N. C. Perry, P. G. Grootenhius, and C. A. G. Haasnoot, *Tetrahedron Comp. Meth.* **3** (1990) 595–613.
13. D. B. Turner, P. Willet, A. Ferguson, and T. W. Heritage, *SAR and QSAR in Environ. Res.* **3** (1995) 101–130.
14. D. A. Thorner, D. J. Wild, P. Willet, and P. M. Wright, *J. Chem. Inf. Comput. Sci* . **36** (1996) 900–908.
15. C. L. Wilkins and M. Randić, *Theor. Chim. Acta* **58** (1979) 45–68.
16. R. Carbo, L. Leyda, and M. Arnau, *Int. J. Quant. Chem.* **17** (1980) 1185–1189.
17. E. E. Hodgkin and W. G. Richards, *Chem. Brit.* **24** (1988) 1141–1144.
18. F. Manuat, F. Sanz, J. Jose, and M. Milesi, *J. Computer-Aided Mol. Design* **5** (1991) 371–380.
19. F. Sanz, F. Manuat, J. Rodriguez, E. Lozoya, and E. Loprezdebrinas, *J. Computer-Aided Mol. Design* **7** (1993) 337–347.
20. A. C. Good and W. G. Richards, *J. Chem. Inf. Comput. Sci*. **32** (1992) 188–191.
21. R. Carbo and L. Domingo, *Int. J. Quant. Chem.* **32** (1987) 517–545.
22. P. E. Bowen-Jenkins and W. G. Richards, *J. Phys. Chem.* **89** (1985) 2195–2197.
23. P. Gaillard, P. Carrupt, B. Testa, and A. Boudon, *J. Computer-Aided Mol. Design* **8** (1995) 83–96.
24. P. M. Dean, *Molecular Foundations of Drug Receptor Interactions*, University Press, Cambridge, 1987.
25. M. Martin, F. Sanz, M. Campillo, L. Prado, J. Perez, and J. Trumo, *Int. J. Quantum Chem.* **23** (1983) 1627–1641 .
26. P. G. Mezey, J. S. Jadav, M. A. Hermsmeier, and T. M. Gund, *J. Mol. Graphics* **6** (1988) 45–53.
27. D. Hadži, J. Koller, and M. Hodošček, *QSAR: Quantitative Structure-Activity Relationships in Drug Design*, Alan R. Liss, Inc., 1989, pp. 259–263.
28. SPARTAN version 4.0, Wavefunction, Inc., 18401 Von Karman Ave. #370, Irvine, CA 927155 U.S.A., @ 1995 Wavefunction, Inc.
29. M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. P. P. Stewart, *J. Am. Chem. Soc.* **107** (1985) 3902–3909.
30. B. H. Besler, K. M. Merz, and P. A. Kollman, *J. Comput. Chem.* **11** (1990) 431–439.
31. A. C. Good and W. G. Richards, *J. Chem. Inf. Comp. Sci.* **33** (1993) 112–116.
32. J. A. Nedler and R. Mead, *Comput. J.* **7** (1965) 308–313.
33. SimMol ver. 1.0 – Fortran 77 program for the molecular similarty calculations made by our group, on the basis of the molecular electrostatic potential and an approximative electron density.
34. J. D. Petke, *J. Comput. Chem.* **14** (1993) 928–933.

SAŽETAK

Usporedni studij indeksa elektrostatske sličnosti

Črtomir Podlipnik i Jože Koller

U ovom radu je molekula etanola superponirana na odabrani skup molekula (etan, etantiol, etilamin, metanol, propan i vodu). Izračunana je sličnost između etanola i molekula u skupu, koja se temelji na molekulskom elektrostatskom potencijalu (MEP, s pomoću AM1 i HF/6-31+G*. Upotrijebljen je Carbov indeks kao mjera sličnosti, a za procjenu sličnosti upotrijebljena su tri različita pristupa. To su MEP s metodom rešetke, ESP naboj s metodom rešetke i ESP naboj s Gaussovom aproksimacijom. Nađeno je da u slučaju velikih Carbovih indeksa izbor molekulsko-orbitalne metode i pristupa za procjenu sličnosti nije odlučujući. U slučaju kada se superponiraju elektrostatski potencijali dviju molekula koje nisu osobtio slične, kao npr. etanol i etan, tada Gaussova aproksimacije daje bjelodano veće vrijednosti indeksa sličnosti od metode rešetke.