FAST EVALUATION OF MOLECULAR 3D SHAPE SIMILARITY#

Črtomir Podlipnik and Jože Koller
Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, 1000 Ljubljana, Slovenia.

#Dedicated to professor Davorin Dolar on occasion of his 80^{th} birthday

Received 14-05-2001

Abstract

In this study the different methods for analytical evaluation of molecular shape similarity are compared. In the first approach, the three Gaussian function approximation was used for description of atom centered electronic density (Good). In the second method the concept of a "hard-sphere" volume was replaced by a soft Gaussian representation (GP). The obtained results were compared with those produced by the single point numerical (grid) shape similarity calculations. It is clear from our study that the general behavior of the GP approximation closely matches that one of the grid-based for fixed orientation calculations. It is also observed that the values for similarity indices of Good method are significantly higher than those obtained with single point numerical grid calculation.

Introduction

Molecular similarity is a fast-emerging concept for material and drug design. Most research efforts so far were concentrated on the development of methods for expressing molecular similarity.[1,2] In this work we concentrated on the evaluation of molecular shape similarity. The molecular shape and quantities derived from it can be used to interpret some features of molecular physical properties. Simple regression models use molecular surface and volume to predict solution properties such as solubility, partitition coefficient, boiling point and rate constant. The shape of molecule plays an important role in drug design. Shape comparison can be helpful for identification of molecules that possess similar spatial (steric) properties but belong to different structural classes. By applying the similarity value as a maximisable function in molecular aligment optimisations, the similarity index can be used to determine the binding models required for 3D QSAR studies and database searches. The 3D structure representations are very important aspects in similarity index calculation and they range from various graph-theoretical, electronic and

quantum-chemical descriptors[3] to various universal 3D molecular structure representation[4,5] and 2D mapping using Kohonen networks.[6]

The measurement of molecular shape similarity through the use of volume overlap was one of the first application of quantitative similarity measure. Hopfinger[7] simply measured the sum of the overlaps between all pairs of atoms in the molecules being compared, using the hard sphere approximation for each atom. Hermann and Herron[8] developed the program OVID, which measures and optimizes the overlap volume between ligand atoms. Masek et al.[9] use for shape comparison more accurate analytical volume calculations presented by Connolly.[10] Good and Richards[11] proposed the use of atom-centered Gaussian functions to approximate electron density. Analytical nature and ease of calculation make this method robust and rapid for similarity optimizations.

Grant and Pickup[12] presented another simple description of molecular shape in which intersecting hard spheres were replaced by the overlapping atom-centered Gaussians. They showed that this approximation is very suitable for shape similarity calculations.[13]

Methodology

CACTVS[14] molecular structure editor (CSED) was used to construct SMILES code[15] of molecular structures. CORINA[16] molecular builder was then applied to build the 3D molecular structures. In our work we used three different similarity indices (Carbo,[17] Hodgkin,[18] Petke[19]) as a quantitative descriptor of the molecular shape. The most widely used form of similarity index applied to calculation of 3D molecular similarity was proposed by Carbo:

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

The numerator in the equation 1 measures property overlap while denominator normalizes similarity result. The difference between equations for Carbo and Petke (Hodgkin) index is only in the denominator part. The sensitivity of the proposed indices to the property magnitude increases in the following order: Carbo < Hodgkin < Petke. The most common procedure for calculation of the similarity index is the grid method described elsewhere.[1,2] We used an alternative approach in which the property distribution is approximated by the sum of the Gaussian functions that can be easily processed analytically. Such analytical evaluations are some orders of magnitude faster (with only minimal effect on its accuracy) than the equivalent numerical calculations.

For the calculation of the molecular shape similarity we have studied and compared two different analytical approaches. In the first approach proposed by Good the electron densities are approximated by the sum of the three atom centered Gaussians. The overlap integrals in the equation are then easily calculated analytically. The equations and parameters for this study were taken from the literature.[11] In the second approach we used a simple description of molecular shape in which intersecting hard spheres were replaced by the overlapping atom centered Gaussians.[12] The atom centered Gaussian density is defined by equation 2 :

$$\rho_i(r_i) = p_i \exp\left(-\alpha_i r_i^2\right) \tag{2}$$

It has been shown[12] that in the equation 2 it is convenient to replace exponent α_i controlling the rate of decay by the scaled exponent:

$$\alpha_i = \frac{\kappa_i}{\sigma_i^2} \tag{3}$$

where σ_i is a "sphere radius" and κ_i is a dimensionless constant. The atomic Gaussian volume V_i^g is deffined by equation 4:

$$V_i^g = \int \rho_i(r_i) d\mathbf{r}_i = p_i(\frac{\pi}{\alpha_i})^{3/2}. \tag{4}$$

The hard sphere atomic volume V_i^{hs} is:

$$V_i^{hs} = \frac{4\pi\sigma_i^3}{3}. \tag{5}$$

We introduced the new parameter λ_i defined by the equation 6:

$$\lambda_i = (\frac{\pi}{\kappa})^{3/2}. \tag{6}$$

It can be shown that if we normalize atomic Gaussian volume with atomic hard-sphere volume then the parameters p_i and λ_i are subject to the following condition:

$$p_i\lambda_i = \frac{4\pi}{3}. \tag{7}$$

Parameters p_i and λ_i reproduce hard-sphere volume independently of the sphere radius. It has been shown[12] that optimal values of atom radius independent parameters are $p = 2.50$ and $\lambda = 1.6755$. The atomic radii used in our calculation were $\sigma(H) = 1.2\mathring{A}$ and $\sigma(C) = 1.70\mathring{A}$.

The Gaussian density of molecular system can be calculated using the equation 8[12]:

$$\rho(\mathbf{r}) = \sum_i \rho_i - \sum_{i<j} \rho_i\rho_j + \sum_{i<j<k} \rho_i\rho_j\rho_k - \dots \tag{8}$$

while the Gaussian volume of molecule is obtained by integration:

$$V^g = \int \rho(\mathbf{r}) d\mathbf{r}. \tag{9}$$

Two orders of the approximation for the overlap Gaussian volumes were used. The first order overlap volume (V^{GP1})(eq. 10) was calculated by retaining the first term in the Gaussian series expansion of density (eq. 8).

$$V^{GP1} \approx \sum_i \sum_j \int \rho_i^A \rho_j^B d\mathbf{r}. \tag{10}$$

The second order overlap volume (V^{GP2})(eq. 11) was calculated with using the first two terms of the Gaussian density expansion (eq. 8).

$$\begin{aligned} V^{GP2} \approx & \sum_i \sum_k \int \rho_i^A \rho_k^B d\mathbf{r} - \\ & \left[\sum_{i<j} \sum_k \int \rho_i^A \rho_j^A \rho_k^B d\mathbf{r} + \sum_i \sum_{k<l} \int \rho_i^A \rho_k^B \rho_l^B d\mathbf{r} \right] + \\ & \sum_{i<j} \sum_{k<l} \int \rho_i^A \rho_j^A \rho_k^B \rho_l^B d\mathbf{r} \end{aligned} \tag{11}$$

Shape similarity index can be easily obtained (eq. 1) from the overlap volumes.[13] The similarity calculations were made by SimMol ver 1.0 program.[21] Simplex method[22] was used for the optimization of similarity indices. The optimization started from the ten randomly generated simplexes. As the result we chose the maximum value of the similarity index obtained from the simplex optimizations. The resulted relative orientations of the pair of molecules was then used for single point numeric (grid method) similarity calculations. Rectangular grid with 2 $\AA$ extent and 0.1 $\AA$ increment was used in this type of calculations.

Results and discussion

Numerical (grid) evaluation of shape similarity indices is time consuming process; however, the theory behind this calculation is less approximate than all the other calculations presented in this paper. This is the reason why we have used results of this type of the calculation as a reference when we compared different analitycal approaches of the similarity index evaluation. In our investigation toluene was compared with a number of benzene derivates. Table 1 lists the results determined from shape similarity calculations.

	GP1		**GP2**		**Good**	
	R_{AB}	H_{AB}	R_{AB}	H_{AB}	R_{AB}	H_{AB}
benzene	**0.913**	**0.908**	**0.904**	**0.900**	**0.936**	**0.933**
	0.900	0.896	0.906	0.902	0.881	0.877
ethyl-	**0.927**	**0.924**	**0.916**	**0.912**	**0.945**	**0.944**
benzene	0.915	0.912	0.916	0.913	0.897	0.895
propyl-	**0.860**	**0.849**	**0.853**	**0.843**	**0.875**	**0.867**
benzene	0.857	0.848	0.857	0.848	0.837	0.829
1-methylethyl-	**0.842**	**0.831**	**0.839**	**0.781**	**0.896**	**0.890**
benzene	0.847	0.839	0.848	0.839	0.839	0.830
butyl-	**0.806**	**0.787**	**0.801**	**0.781**	**0.821**	**0.807**
benzene	0.807	0.791	0.807	0.791	0.793	0.777
2-methylpropyl-	**0.805**	**0.785**	**0.794**	**0.775**	**0.818**	**0.807**
benzene	0.811	0.794	0.811	0.795	0.789	0.773
1-methylpropyl-	**0.788**	**0.767**	**0.788**	**0.769**	**0.839**	**0.827**
benzene	0.802	0.793	0.800	0.784	0.792	0.776
1,1-dimethylethyl-	**0.792**	**0.772**	**0.798**	**0.779**	**0.858**	**0.848**
benzene	0.807	0.793	0.807	0.791	0.808	0.792

Table 1: Similarity results for shape comparison of toluene with all other compounds listed in the table. Results obtained from analytical methods are bolded while numerical results are presented with normal text. Legend: **GP1** (**GP2**) - The first (second) order Grant-Pickup approximation of Gaussian volumes; **Good** - Three Gaussian approximation of electronic density; R_{AB} - Carbo index; H_{AB} - Hodgkin index.

We found that the shape similarity indices obtained with approaches GP1 and GP2 did not differ significantly from all the comparisons we have made. The values of similarity indices calculated with Good method are mainly larger than the ones obtained with GP1 and GP2 approaches. For all three approaches, the obtained results were compared with those produced by the single point numerical (grid) shape similarity calculations. It is clear from our study that the general behavior

of the GP1 (GP2) approximation closely matches that one of the grid-based for fixed orientation calculations. It is also observed that the values of shape similarity index of Good method are significantly higher than those obtained with single point numerical grid calculation. Similar trend was observed in previous study of electrostatic similarity[23] where the numerical grid method[1] and anaytical Good method are compared. Table 2 shows the details of similarity results for shape comparison of toluene and 1,1-dimethylethylbenzene.

	GP1	**GP2**	**GOOD**	**GRID**[a]
min R_{AB}	0.692	0.703	0.854	-
max R_{AB}	0.792	0.798	0.858	0.807
av. R_{AB}	0.769	0.787	0.858	-
σ	0.042	0.029	0.001	-
iterations [b]	1101	1280	1168	1
CPU time [c]	1	10	7	8500

Table 2: Similarity results for shape comparison of toluene and 1,1-dimethylethylbenzene. σ - Standard deviation of optimized similarity index. R_{AB} - Carbo index. [a] Single point similarity calculation. [b] Total number of iterations (Simplex optimisation is started from ten different random simplexes). [c] Relative CPU time per iteration.

These results show that speed increases up to 3 orders of magnitude when analytical functions are used in place of grid-based shape similarity calculation. It has been also shown that the GP1 (GP2) approaches are more sensitive due to choice of the initial relative orientation for simplex optimization than Good approach.

Conclusions

In our work we tested different approaches for fast evaluation of shape similarity index. We found that both approaches based on Grant-Pickup (GP) approximation produce similar results. Approach based on first order GP approximation is much faster than that based on the second order approximation. Therefore, the first one is preffered for fast and robust molecular similarity calculation. The great advantage of the GP approximations is customization of atomic Gaussian density via two parameters p and λ (eqs. 6 and 7) that reproduce the hard-sphere volume independently of sphere radius σ_i. More diffuse atomic centered Gaussian function for description of hydrogen bonding area in molecule may be used. Evaluation of shape similarity using GP aproximation is fast and robust and may be used in drug design for rational selection of candidates from large databases.

References

1. G. M. Maggoira and M.A. Johnson (Eds.), *Concepts and Application of Molecular Similarity*, John Willey and Sons, 1990.
2. P.M. Dean (Ed.), *Molecular Similarity in Drug Design*, Blackie Academic & Professional, 1995.
3. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors, in Series of Methods and Principles of Medicinal Chemistry - vol 11*, Ed. R. Mannhold, H. Kubinyi and H. Timmerman. Willey - VCH, Weinheim, 2000.
4. M. Novic and J. Zupan, *A New General and Uniform Structure Representation, in Software Development in Chemistry - vol 10*, Ed. J. Gasteiger, GDCh, Frankfurt am Main, 1996, pp. 47-58.
5. S. Bauerschmidt and J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 705-714.
6. J. Polanski, J. Gasteiger, M. Wagner and J. Sadowski, *Quant. Struct.-Act. Relat.* **1998**, *17*, 27-36.
7. A.J. Hopfinger, *J. Am. Chem. Soc.* **1980**, *102*, 7196-7206.
8. R.B. Hermann and D.K. Herron, *J. Comput. Aid. Mol. Des.* **1991**, *5*, 511-524.
9. B.B. Masek, A. Merchant, and J.B. Matthew, *J. Med. Chem.* **1993**, *36*, 1230-1238.
10. M.L. Connoly, *J. Am. Chem. Soc.* **1985** *107*, 5959-5967.
11. A.C. Good, and W.G. Richards, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112-116.
12. J.A. Grant and B.T. Pickup, *J. Phys. Chem.* **1995**, *99*, 3503-3510.
13. J.A. Grant, M.A. Gallardo and B.T. Pickup, *J. Comput. Chem.* **1996**, *17*, 1653-1666.
14. W.D. Ihlenfeldt, Y. Takahashi, H. Abe and S. Sasaki, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109-116.
15. D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31-36.
16. J. Sadowski, J. Gasteiger and G. Klebe, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000-1008.
17. R. Carbo, L. Leyda and M. Arnau, *Int. J. Quant. Chem.* **1980**, *17*, 1185-1189.
18. E.E. Hodgkin and W.G. Richards, *Int. J. Quant. Chem. Quantum Biology Simposia* **1987**, *14*, 105-110.
19. J.D. Petke, *J. Comput. Chem.* **1993**, *14*, 928-933.
20. Good, A.C., *Molecular Similarity and Drug Design, edited by Dean, P.M.*, Blackie Academic & Professional, 1995, pp 24-56.
21. SimMol ver 1.0 - Fortran 77 program for molecular similarity calculation made by our group.
22. J.A. Nedler and R. Mead, *Comput. J.* **1965**, *7*, 308-313.
23. C. Podlipnik and J. Koller, *Croat. Chem. Acta* **1998**, *71*, 689-696.

Povzetek

V študiji smo primerjali različne metode za hiter analitični izračun molekulske podobnosti na osnovi njihove oblike. Pri prvem pristopu smo za opis oblike molekule uporabili linearno kombinacijo treh Gaussovih funkcij. Pri drugem pristopu pa smo volumne atomov v molekuli namesto z modelom toge krogle opisali z Gaussovo funkcijo (GP). Dobljene rezultate smo primerjali s tistimi dobljenimi s pomočjo numerične mrežne metode. Iz naše študije je razvidno, da imajo indeksi podobnosti, dobljeni s pomočjo GP aproksimacij, zelo podobne vrednosti kot jih da numerična mrežna metoda. Prav tako je ugotovljeno, da imajo indeksi podobnosti, dobljeni s pomočjo Goodove metode, precej višje vrednosti kot tisti, dobljeni kot rezultat numeričnega mrežnega izračuna.