*MATCH*
*Communications in Mathematical*
*and in Computer Chemistry*

**no. 52, September 2004**
**MATCDY (52) 55-63(2004)**

**ISSN 0340 - 6253**

# A new method for indirect evaluation of molecular shape similarity

Črtomir Podlipnik[a] , Tomaž Šolmajer[b] and Jože Koller[a]

[a] Faculty of Chemistry and Chemical Technology, University of Ljubljana.
Aškerčeva 5, SI-1000 Ljubljana, Slovenia

[b] National Institute of Chemistry, Hajdrihova 19, SI-1000 Ljubljana, Slovenia

Received February 25, 2004

**Abstract**

In this paper a new method for indirect evaluation of molecular shape similarity is introduced. In the first step of molecular comparison a conversion of molecular 3D-structure to translational and rotational invariant RDF code is performed. Secondly, the similarity indices are computed based on the RDF code comparison for each pair of molecules. These similarity indices are then used as descriptors for generating QSAR/QSPR models. In a practical example we have used the approach to correlate the octane isomers structure and their octane number. The result compares favorably with models obtained by the use of various topological indices.

## 1 Introduction

Molecular similarity is an important tool in material and drug design [1,2]. It aims to give a quantitative answer to the question of how similar two given molecules are. Molecular similarity theory can be used in a variety of topics: as an indicator of molecular chirality [3]; as a functional for finding optimized molecular alignments [4]; to compare different theoretical calculation methods [5]; and as molecular descriptors to build quantitative structure-activity/property relationships (QSAR/QSPR) [6-8]. Similarity index is a quantitative measure of similarity between two molecules. However, in general similarity index depends on relative orientation of two molecules. Thus, in order to find the maximum value of similarity index various optimization techniques [4,9–12] have been applied to solve this problem.

In this work an indirect method for evaluation of molecular similarity is introduced. In the first step of the method the conversion of molecular 3D-structures to translational and rotational invariant RDF (Radial Distribution Function) codes is applied. Gasteiger et al. [14] proposed radial distribution function (RDF) as a new 3D-molecular descriptor. This function is well documented in physics and physical chemistry in general and in X-ray diffraction in particular [15]. In the second step of the procedure similarity indices for each pair of molecules are calculated by comparison of their respective RDF codes. The results of similarity calculations are collected in a similarity matrix. The RDF similarity indices are subsequently used as descriptors for generation of simple regression models.

## 2 Methodology

The RDF code of a molecule is calculated by the following equation:

$$G(R) = \sum_{j=1}^{n} \sum_{i=j+1}^{n} p_i p_j \exp\left(-B(R - r_{ij})^2\right), \tag{1}$$

where $n$ is number of atoms, $p_i$ and $p_j$ are properties of $i$-th and $j$-th atom, $r_{ij}$ is the distance between the atoms $i$ and $j$, $B$ is a smoothing parameter, which defines the probability distribution of the individual interatomic distance. RDF code is usually calculated at a number of discrete points $R$ within selected interval. RDF codes have some important and highly useful properties such as translational and rotational invariance of the entire molecule. the length of the code is independent of the size of a molecule, the code is unambiguous regarding to 3D arrangement of atoms. The molecular geometries needed for these calculations have been generated by using the 3D Structure Generator CORINA [16] that is accessible to the scientific community via internet.

In our work we used three different similarity indices (Carbo [17], Hodgkin [18], Petke [19]) to quantify RDF similarity. The most widely used form was proposed by Carbo.

$$C_{AB} = \frac{\int P_A P_B dV}{(\int P_A^2 dV)^{1/2} (\int P_A^2 dV)^{1/2}} \tag{2}$$

The numerator in equation 2 measures property overlap while denominator normalizes similarity result.

The Carbo index, quantitative measure for similarity between RDF codes of molecules A and B is then calculated by equation:

$$C_{AB} = \frac{\sum_{k=1}^{M} G_A(R_k) G_B(R_k)}{(\sum_{k=1}^{M} G_A^2(R_k))^{1/2} (\sum_{k=1}^{M} G_B^2(R_k))^{1/2}}, \tag{3}$$

where $M$ is number of points that represent RDF code, $G_A$ and $G_B$ are RDF codes of molecules A and B, respectively, and $R_k$ is a discrete point within defined interval. Due to the invariant properties of RDF code. the value of similarity index is independent of relative intermolecular orientation. A simple

C++ program using Openbabel open-source library [20] has been written for calculation and comparison of RDF codes.

Program SimMol [21] has been subsequently used for calculation of 3D molecular shape similarity. The shape of molecules is described with their approximative electron density [22] and Simplex algorithm [9] has been applied to maximize molecular similarity indices.

## 3 Results and Discussion

Firstly, in Section A we documented the optimization of parameters used in computation of similarity indices: RDF distance interval and smoothing parameter. As an application of this methodology motor octane number (MON) for a series of octane isomers were computed (Section B).

**Section A**

The shape of RDF code depends on several variables such as: step size; $B$ - value of smoothing parameter ; $p_i$ - characteristic properties of atoms or fragments. The atomic numbers of non-hydrogen atoms are used as characteristic properties in our case. Table 1 lists the values of various similarity indices between octane and 2,2-dimetylhexane obtained by comparison of RDF codes. RDF is defined on interval between 0-12 Å and smoothing parameter is set to 25 Å$^{-2}$. The RDF has been calculated on equaly distributed points within defined interval that has been divided in to 4-256 subintervals.

Table 1: The dependance of values of similarity indices between octane and 2,2-dimethylhexane on number of points equally distributed in interval from 0 Å to 12 Å. Smoothing parameter $B$ is set to 25 Å$^{-2}$.

| $M$ | Carbo | Hodgkin | Petke |
|---|---|---|---|
| 256 | 0.8894 | 0.8866 | 0.8208 |
| 128 | 0.8894 | 0.8866 | 0.8208 |
| 64 | 0.8894 | 0.8866 | 0.8207 |
| 32 | 0.8879 | 0.8871 | 0.8521 |
| 16 | 0.8998 | 0.8996 | 0.8818 |
| 8 | 0.9198 | 0.9198 | 0.9191 |
| 4 | 0.0594 | 0.0270 | 0.0143 |

We found that convergence of similarity indices values is obtained if the number of subintervals exceeds 64. The maximum of similarity indices has been found at $M = 8$, but these results are not relevant due to infrequent sampling.

The effect of the smoothing parameter on the values of the similarity indices has also been studied. The RDF has been calculated at 129 equally distributed points in interval from 0 Å to 12 Å and smoothing factor varied from 1 $Å^{-2}$ to 250 $Å^{-2}$. The results of this study are collected in table 2.

Table 2: The smoothing parameter dependance of similarity indices values between octane and 2,2-dimethylhexane. The RDF is defined at 129 equally distributed points in interval from 0 Å to 12 Å.

| $B$ [$Å^{-2}$] | Carbo | Hodgkin | Petke |
|---|---|---|---|
| 1.0 | 0.9626 | 0.9491 | 0.8133 |
| 2.5 | 0.9467 | 0.9316 | 0.7905 |
| 10.0 | 0.9049 | 0.8974 | 0.7948 |
| 25.0 | 0.8894 | 0.8866 | 0.8208 |
| 100.0 | 0.8748 | 0.8735 | 0.8279 |
| 250.0 | 0.8625 | 0.8621 | 0.8375 |

The shape of RDF function depending on smoothing parameter of 2,2 dimethylhexane is illustrated on Figure 1.

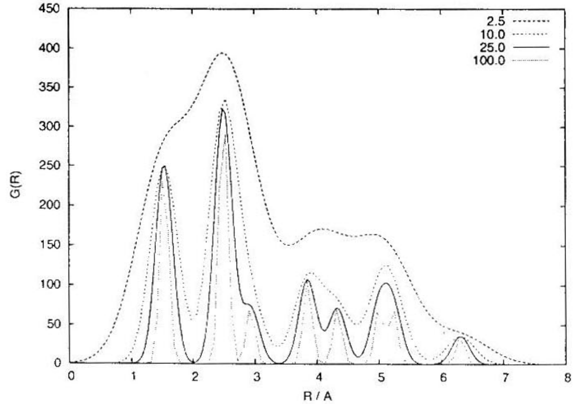

Figure 1: The shape of RDF function of 2,2 dimethylhexane. Variation of smoothing factor.

It can be observed that the shape of RDF function (Fig. 1) strongly depends on smoothing parameter; higher values of smoothing parameter lead to more structured RDF functions. The high values of similarity indices have been obtained when we compared two unstructured RDF functions (low values of smoothing parameter). Comparison of two very structured RDF functions (high values of smoothing parameter) lead to lower similarity indices than should be expected. According to our study the reasonable values of step size and smoothing parameter for RDF similarity calculation should be 0.1 Å and 25 $Å^{-2}$, respectively. The sensitivity of similarity indices to the property magnitude increases in the following order: Carbo < Hodgkin < Petke.

Section B

As an example of application of our approach we used similarity matrices to describe a set of octane isomers in order to search for QSPR models correlating the octane number of the molecules with their similarity to each other.

Motor octane number (MON) is figure of merit representing the resistance of gasoline to premature detonation when exposed to heat and pressure ín the combustion chamber of an internal-combustion engine. The values of selected RDF similarity indices and motor octane number for octane isomers are listed in table 3.

Table 3: The values of selected RDF similarity indices and motor octane number for octane isomeres.

| No. | octane | ${}^{r}C_8$ | ${}^{r}C_9$ | ${}^{r}C_{14}$ | ${}^{r}C_{18}^{*}$ | MON [23] |
|---|---|---|---|---|---|---|
| 1 | n-octane | 0.8717 | 0.8894 | 0.7891 | 0.7427 | - |
| 2 | 2-methylheptane | 0.9418 | 0.9644 | 0.8849 | 0.8381 | 23.8 |
| 3 | 3-methylheptane | 0.9644 | 0.9740 | 0.9141 | 0.8700 | 35.0 |
| 4 | 4-methylhexane | 0.9746 | 0.9706 | 0.9261 | 0.8728 | 39.0 |
| 5 | 2,3-dimethylhexane | 0.9938 | 0.9885 | 0.9681 | 0.9294 | 78.9 |
| 6 | 2,4-dimethylhexane | 0.9896 | 0.9887 | 0.9695 | 0.9294 | 69.9 |
| 7 | 2,5-dimethylhexane | 0.9684 | 0.9947 | 0.9416 | 0.8910 | 55.7 |
| 8 | 3,4-dimethylhexane | 1.0000 | 0.9705 | 0.9820 | 0.9461 | 81.7 |
| 9 | 2,2-dimethylhexane | 0.9705 | 1.0000 | 0.9528 | 0.9136 | 77.4 |
| 10 | 3,3-dimethylhexane | 0.9945 | 0.9764 | 0.9912 | 0.9627 | 83.4 |
| 11 | 3-ethylhexane | 0.9710 | 0.9538 | 0.9462 | 0.8822 | 52.4 |
| 12 | 2,3,4-trimethylpentane | 0.9886 | 0.9515 | 0.9950 | 0.9730 | 95.9 |
| 13 | 2,3,3-trimethylpentane | 0.9729 | 0.9343 | 0.9954 | 0.9914 | 99.4 |
| 14 | 2,2,3-trimethylpentane | 0.9820 | 0.9528 | 1.0000 | 0.9839 | 99.9 |
| 15 | 2,2,4-trimethylpentane | 0.9774 | 0.9649 | 0.9890 | 0.9660 | 100.0 |
| 16 | 3-ethyl-2-methylpentane | 0.9696 | 0.9282 | 0.9719 | 0.9571 | 88.1 |
| 17 | 3-ethyl-2-methylpentane | 0.9603 | 0.9132 | 0.9750 | 0.9769 | 88.7 |
| 18 | 2,2,3.3-tetramethylbutane | 0.9136 | 0.9461 | 0.9839 | 1.0000 | - |

* - The similarity index ${}^{r}C_{18}$, for example. is Carbo index of RDF similarity between molecule 18 and a molecule listed in table 3.

The elements of the RDF similarity matrix have been imported into the program CODESSA [24] as descriptors. The multiparameter regression models have been then generated using a heuristic parameter selection method as given within CODESSA.

Equations from 4 to 6 represent one to three parameter regression models for correlation between RDF similarity of octanes and their octane numbers.

$$\begin{aligned} MON &= -398(\pm 32) + 507(\pm 34)\,{}^{r}C_{18} \\ &(n = 16,\ R = 0.970,\ s = 6.23,\ F = 221.85,\ Q = 0.962); \end{aligned} \tag{4}$$

$$\begin{aligned} MON &= -335(\pm 99) + 851(\pm 60)\,{}^{r}C_{14} - 420(\pm 133)\,{}^{r}C_{8} \\ &(n = 16,\ R = 0.978,\ s = 5.6,\ F = 139.45,\ Q = 0.969); \end{aligned} \tag{5}$$

$$\begin{aligned} MON &= -352(\pm 99) + 922(\pm 83)\,{}^{r}C_{14} - 577(\pm 183)\,{}^{r}C_{8} + 104(\pm 85)\,{}^{r}C_{9} \\ &(n = 16,\ R = 0.980,\ s = 5.49,\ F = 97.09,\ Q = 0.965). \end{aligned} \tag{6}$$

It can be observed from statistical results attached to equations from 4 to 6 that adding of new parameters have no significant effect to improvement of correlation between RDF similarity indices and MON. The goodness-of-prediction has been measured with leave-one-out cross-validation coefficient $Q$. The high values of cross-validation coefficient $Q$ show good predictive ability for all three models.

Figure 2 shows the correlation between experimental and calculated values of MON for a set of octane isomers. The calculated values were computed by equation 6.

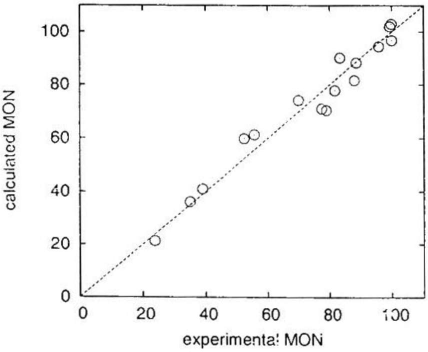

Figure 2: Correlation between experimental and calculated value of octane number. The calculated values were computed by equation 6.

Results of one parameter regressions obtained from our similarity matrices approach can be compared to corresponding one parameter regressions from the literature [25].

Table 4: The regression statistics for eight mono-parametric linear structure-motor octane models for octanes. Legend: $R$ - correlation coefficient, $s$ - standard error of estimate and $F$ - Fisher ratio.

| descriptor | | $R$ | $s$ | F | ref. |
|---|---|---|---|---|---|
| $\chi$ | Randić index | 0.778 | 16.0 | 21.5 | [25] |
| $\epsilon$ | edge-connectivity index | 0.271 | ·24.6 | 1.1 | [25] |
| $J$ | Balaban index | 0.928 | 9.5 | 86.4 | [25] |
| $^{c}J$ | reversed Balaban index | 0.966 | 6.6 | 193.1 | [25] |
| $^{r}J$ | Harary-Balaban index | 0.963 | 6.9 | 180.6 | [25] |
| $J'$ | quotient Balaban index $1^{st}$ kind | 0.921 | 10.0 | 178.0 | [25] |
| $J''$ | quotient Balaban index $2^{nd}$ kind | 0.965 | 6.7 | 190.0 | [25] |
| $^{s}C_{18}$ | Carbo index (shape similarity) | 0.969 | 6.4 | 210.9 | |
| $^{r}C_{18}$ | Carbo index (RDF similarity) | 0.970 | 6.2 | 221.5 | |

It can be seen from the results collected in table 4 that the quality of one parameter correlations depends on ability of a descriptor to describe branching of octane isomers. According to this observation we may conclude that RDF similarity matrices are useful for generating regression models that describe correlation between molecular similarity and a molecular property such as its octane number. The one parameter correlations obtained from RDF similarity matrices are compared favorably to ones obtained by the use of topological indices. A good regression model between shape (approximative electron density) similarity index - $^{s}C_{18}$ and octane numbers has also been found.

## 4 Conclusions

The important feature of RDF code is rotational and translational invariance, therefore optimization of RDF code alignment is not needed. The RDF similarity matrices are successfully used as descriptors for correlation between octane isomers structure and their octane number. It seems that the quality of one parameter correlations between calculated and experimental values of octane number depends on ability of descriptor to describe branching of octane isomers. The use of RDF molecular similarity indices as descriptors is a promising tool for generating QSAR/QSPR models.

## References

[1] P. M. Dean. (ed.), Molecular Similarity in Drug Design, Blackie Academic Professional, 1995.

[2] R. Carbo-Dorca, D. Robert, L. Amat, X. Girones, E. Besalu, Molecular Quantum Similarity in QSAR and Drug Design, Springer-Verlag Berlin Heidelberg New York, 2000.

[3] P. Mezey, R. Ponec, L. Amat, R. Carbo-Dorca, Quantum similarity approach to the charecterization of molecular chirality., *Enantiomeres* 4 (1999) 371-378.

[4] P. Contans, L. Amat, R. Carbo-Dorca, Toward a global maximization of the molecular similarity function: Superposition of two molecules., *J. Comput. Chem.* 18 (1997) 826–846.

[5] M. Sola, J. Mestres, R. Carbo, M. Duran, A comparative analysis by means of quantum molecular similarity measures of density distributions derived from conventional ab initio and density functional methods., *J. Chem. Phys.* 104 (1996) 636–647.

[6] A. C. Good, S. S. So, W. G. Richards, Structure-activity relationships from molecular similarity matrices., *J. Med. Chem.* 36 (1993) 433–438.

[7] S. S. So, M. Karplus, Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. method and validations., *J. Med. Chem.* 40 (1997) 4347–4359.

[8] L. Amat, R. Carbo-Dorca, R. Ponec, Simple linear QSAR models based on quantum similarity measures, *J. Med. Chem.* 42 (2003) 5169–5180.

[9] J. A. Nedler, R. Mead, A simplex method for function minimization, *Comput. J.* 7 (1965) 308–313.

[10] M. F. Parretti, R. T. Kroemer, J. H. Rothman, W. G. Richards, Alignment of molecules by the Monte Carlo cptimization of molecular similarity indices., *J. Comput. Chem.* 18 (1997) 1344–1353.

[11] A. J. McMahon, P. M. King, Optimization of Carbo molecular similarity index using gradient methods., *J. Comput. Chem.* 18 (1997) 151–158.

[12] D. J. Wild, P. Willett, Similarity searching in files of three-dimensional chemical structures: Aligment of molecular electrostatic potentials with genetic aghorithm., *J. Chem. Inf. Comp. Sci.* 36 (1996) 159–167.

[13] M. Hemmer, V. Steinhauer, J. Gasteiger, The prediction of the 3D structure of organic molecules from their infrared spectra, *Vibrat. Spectroscopy* 19 (1999) 151–164.

[14] V. Steinhauer and J. Gasteiger, Obtaining 3D structure from infrared spectra of organic compounds using neural networks, in *Software-Entwicklung in der Chemie 11*, G. Fels, V.Schubert (eds.), Gesellshaft Deutcher Chemiker, Frankfurt/Main, 1997.

[15] J. Karle, Applications of Mathematics to Structural Chemistry, *J. Chem. Inf. Comput. Sci.* 34 (1994) 381–390.

[16] J. Gasteiger, C. Rudolph, J. Sadowski, Automatic generation of 3D-atomic coordinates for organic molecules., *Tetrahedron Comp. Method.* 3 (1990) 537–547.

[17] R. Carbo, L. Leyda, M. Arnau, How similar is a molecule to another? An electron density measure of similarity between two molecular structures., *Int. J. Quant. Chem.* 17 (1980) 1185–1189.

[18] E. E. Hodgkin, W. G. Richards, Molecular similarity., *Chem. Brit.* 24 (1988) 1141–1144.

[19] J. D. Petke, Cumulative and discrete similarity analysis of electrostatic potential and fields. *J. Comput. Chem.* 14 (1993) 928–933.

[20] http://openbabel.sourceforge.net/.

[21] Č. Podlipnik, J. Koller, Fast evaluation of molecular 3D shape similarity, *Acta Chim. Slov*, 48 (2001) 325-331.

[22] A. C. Good, W. G. Richards, Rapid evaluation of shape similarity using Gaussian functions, *J. Chem. Inf. Comp. Sci.* 33 (1993) 112-116.

[23] A. T. Balaban, I. Motoc, Chemical graphs. XXXVI. Correlations between octane numbers and topological indices of alkanes, *MATCH Commun. Math. Comput. Chem.* 5 (1979) 197-218.

[24] Codessa 2.6, Semichem Inc., Sawnee Mission, KS, 2001.

[24] S. Nikolić, D. Plavšić, N. Trinajstić, On the Balaban-like topological indices, *MATCH Commun. Math. Comput. Chem.* 44 (2001) 361-386.