MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **56** (2006) 261-270

ISSN 0340 - 6253

Similarity of radial distribution function's intervals

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 March 20, 2006)

Abstract

We tried to improve descriptive ability of RDFs similarity indices. New similarity descriptors are result of comparision RDFs at different intervals. In a practical example we used RDFs similarity descriptors to correlate the CBG activity of steroids. The models with high internal predictivity measured with Q^2 have been reported in this work. On the other hand applicabilty of this method for QSAR modeling is rather limited due the fact that few descriptors apearing in proposed models are selected from very large pool of RDF similarity descriptors (217 descriptors vs 31 compounds) this may lead to chance correlation and missleading statistics.

1 Introduction

Structure-based descriptors can be related to a relevant molecular property or biological activity by some statistical procedure in manner to derive quantitative structure property/activity relationships (QSPR/QSAR) models. An alternative approach for determining structure property/activity correlations is using an elements from similarity matrices as a molecular descriptors. The rationale for similarity lies in the similarity-property principles which states that structural similar molecules tends to exhibit similar properties.

In this work we generated molecular similarity indices with comparison of selected intervals of RDFs rather than complete codes as we did in our previous work [1] where we used RDF similarity indices in manner to generate models for prediction octane numbers for octane isomers.

The application of method deals with a dataset of 31 steroids that bind to the *corticosteroid binding globuline* (CBG) receptor. This dataset has been formed for the introduction the widely used *Comparative Molecular Field Analysis* by Cramer et al [4]. Molecular similarity calculations [5–8] and autocorrelation of molecular surface properties [9] were performed for these molecules and analyzed by neural networks and statistical methods.

2 Methodology

The molecular geometries of 31 steroids that binds to CBG have been found on website of prof. Gasteiger's research team [10]. Gasteiger atomic charges [11] are calculated with PETRA [12] (Property Estimation of for the Treatment of Reactivity Applications). The charges fitted from electrostatic potential (ESP) [13] and Mulliken charges [14] are calculated with AMPAC [15] molecular modeling package using semiempirical AM1 method [16]. Gasteiger*, ESP* and Mulliken* are atomic charges calculated with summation of hydrogen charges on heavy atom charge.

Gasteiger et al. [2] 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 [3]. The RDF ($G(r)$) of molecule is defined by 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 (p_j) is property of atom i (j), r_{ij} is the distance between atoms i and j, B is a smoothing parameter.

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

$$C_{l,m} = \frac{\int P_l P_m \quad dV}{(\int P_l^2 \quad dV)^{1/2} (\int P_m^2 \quad dV)^{1/2}} \tag{2}$$

The numerator in equation 2 measures property overlap while denominator normalizes similarity result. To increase sensitivity of sensitivity to property magnitude Hodgkin et al. proposed a modification of the Carbo index:

$$H_{l,m} = \frac{2 \int P_l P_m \quad dV}{\int P_l^2 \quad dV + \int P_m^2 \quad dV} \tag{3}$$

We used these two formulations for definition of similarity of partial radial distribution functions.

$$C^k_{l,m} = \frac{\sum\limits_{i\in k} G_{l,i}G_{m,i}}{\left[\sum\limits_{i\in k} G^2_{l,i}\right]^{1/2}\left[\sum\limits_{i\in k} G^2_{m,i}\right]^{1/2}}, \tag{4}$$

$$H^k_{l,m} = \frac{2\sum\limits_{i\in k} G_{l,i}G_{m,i}}{\sum\limits_{i\in k} G^2_{l,i} + \sum\limits_{i\in k} G^2_{m,i}}, \tag{5}$$

$C^k_{l,m}$ ($H^k_{l,m}$) is similarity index between RDF distributions of molecules l and m in k-th interval. RDF is defined in equally distributed points (10 points/Å), smoothing parameter is set to 25 $Å^{-2}$, width of each of 7 intervals is 1.6 Å (the range of the length of C-C bond). C++ program using Openbabel open-source library [19] has been written for calculation and comparison of partial RDF codes. The RDF similarity indices are then collected in similarity matrix. The elements of the RDF similarity matrix have been imported into the program CODESSA [20] as external descriptors. The multiparameter regression models have been then generated using a heuristic parameter selection method as given within CODESSA.

3 Results and Discussion

As an example of application of our approach we used similarity matrices to describe a set of corticosteroids in order to search for QSAR models correlating the CBG activity of the molecules with their RDF similarity to each other. CBG affinity data of 31 steroids are collected in table 1.

Table 1: CBG binding affinity data. Names, CBG activity and activity classes for set of 31 steroids. Activity classes: 1 - high activity; 2 - intermediate activity; 3 - low activity.

No.	name	pK $\log(1/K)$	activity class
1	aldosterone	-6.279	2
2	androstanediol	-5.000	3
3	5-androstenediol	-5.000	3
4	4-androstenedion	-5.763	3
5	androsterone	-5.613	3
6	corticosterone	-7.881	1
7	cortisol	-7.881	1
8	cortisone	-6.892	2
9	dehydroepiandrosterone	-5.000	3
10	11-deoxycorticosterone	-7.653	1
11	11-deoxycortisol	-7.881	1
12	dihydrotestosterone	-5.919	2
13	estradiol	-5.000	3
14	estriol	-5.000	3
15	estrone	-5.000	3
16	etiocholanolone	-5.255	3
17	pregnenolone	-5.255	3
18	17α-hydroxypregnenolone	-5.000	3
19	progesterone	-7.380	1
20	17α-hydroxyprogesterone	-7.740	1
21	testosterone	-6.724	2
22	prednisolone	-7.512	1
23	cortisolacetate	-7.553	1
24	4-pregnene-3,11,20-trione	-6.779	2
25	epicorticosterone	-7.200	1
26	19-nortestosterone	-6.144	2
27	16α,17α-dihydroxyprogesterone	-6.247	2
28	16α-methylprogesterone	-7.120	1
29	19-norprogesterone	-6.817	2
30	2α-methylcortisol	-7.688	1
31	2α-methyl-9α-fluorocortisol	-5.797	2

Figure 1 represents comparison of the RDF of 2α-methylcortisol and 2α-methyl-9α-fluorocortisol at four different intervals. Gasteiger* charges are used as a characteristic atomic property in RDF definition. Results of similarity indices calculations between these two molecules for the four different intervals are collected in table 2.

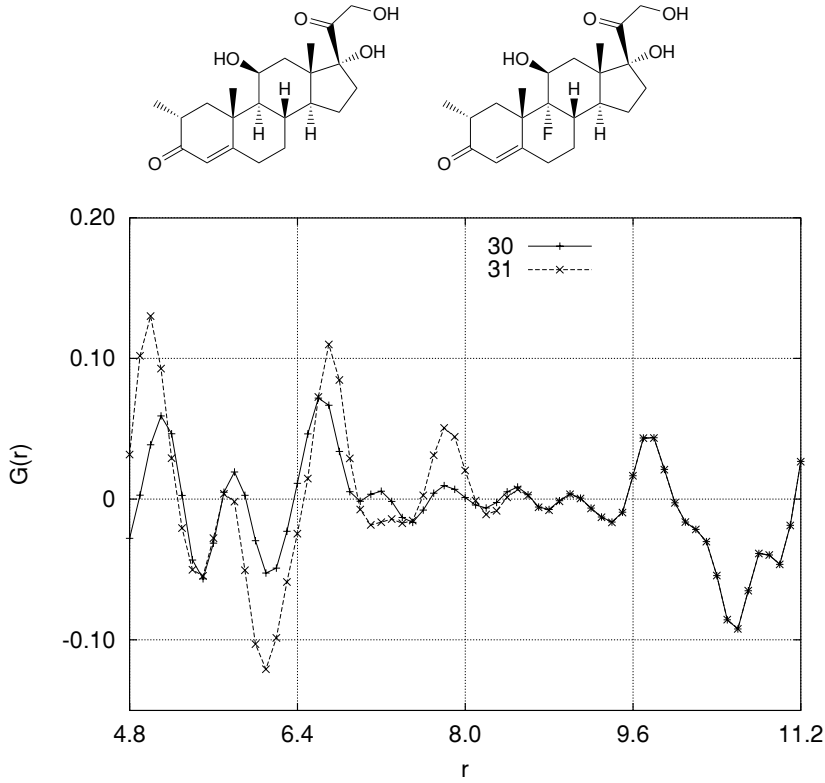


Figure 1: RDF comparison of 2α-methylcortisol and 2α-methyl-9α-fluorocortisol.

From the figure 1 and table 2 it can be observed that the RDF of both molecules are the same at the interval between 9.6 and 11.2 Å and they are the most different at the interval between 4.8 and 6.4 Å. From the results collected in table 2 it can be also seen that the Hodgkin type index is more sensitive than Carbo type index. According to this fact Hodgkin type indices were used in generation of QSAR models.

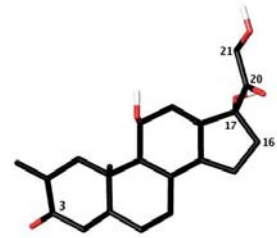
Table 2: Similarity of partial RDF between 2α-methylcortisol and 2α-methyl-9α-fluorocortisol. C^k (H^k) - Carbo (Hodgkin) index of k-interval.

k	interval	C^k	H^k
4	4.8-6.4 Å	0.7738	0.6151
5	6.4-8.0 Å	0.8084	0.7369
6	8.0-9.6 Å	0.8101	0.7863
7	9.6-11.2 Å	1.0000	1.0000

The dependance of the RDF similarity indices on conformational changes of 2α-methylcortisol has been studied. The indices collected in table 3 are measure of similarity between conformation generated with CORINA and conformations listed in table 3 obtained by rotation of side chain at the position C_{17} about the wedge bond.

Table 3: The dependence of the RDF similarity indices at four different intervals on conformational changes of side chain at the position C_{17} of 2α-methylcortisol.

	torsion angle	Similarity index			
conf.	O=C_{20}-C_{17}-C_{16}	H^4	H^5	H^6	H^7
1	-20.8°	1.0000	1.0000	1.0000	1.0000
2	-60°	0.5047	0.9882	0.3557	-0.4300
3	-30°	0.7949	0.9982	0.8729	0.5108
4	0°	0.3608	0.9792	0.4183	0.9367
5	30°	0.8949	0.9804	-0.2950	0.8832
6	60°	0.8299	0.9733	-0.1512	0.8810



From the results collected in table 3 it can be seen that the value of RDF similarity indices depends strongly on a molecule's conformation. Fortunately, steroids taken as an example are rather rigid molecules and some of them have flexibile side chain at the position C_{17}. The orientation of the this side chain was left the same as is generated by CORINA (tipical value for torsion angle O=C_{20}-C_{17}-C_{16} is -20.8°) .

The effect of charge model choice on the quality of the regression models for the correlation between RDF similarity of steroids and their CBG activity has been studied. The results of this study are collected in table 4.

Table 4: Comparison of five parameter regression models for correlation between RDF similarity of steroides and their CBG activity.

Charge	R^2	F	s^2	Q^2
Gasteiger	0.919	56.87	0.12	0.885
Gasteiger*	0.933	70.10	0.09	0.905
Mulliken	0.896	43.11	0.15	0.849
Mulliken*	0.936	72.66	0.09	0.902
ESP	0.910	50.28	0.13	0.874
ESP*	0.905	47.64	0.13	0.864

It can be observed from statistical results collected in table 4 that choice of charge model has no significant effect on quality of regression models. The values of leave one out cross-validation coefficient Q^2 lies between 0.849 (Mulliken) and 0.905 (Gasteiger*). The Gasteiger* charges (hydrogen charges summed on heavy atom) were selected as a characteristic property in RDF definition (equation 1), similarity indices calculation (equation 5) and generation of QSAR models, consequently. The calculation of Gasteiger* charges is straight forward of without solving an electron molecular structures by applying of semiempirical or ab initio methods.

The heuristic method within CODESSA were used for selection of descriptors from the base of 217 (31 compounds x 7 intervals) RDF similarity indices and for generation of QSAR models.

Equations from 6 to 9 represent two to five parameters regression models for correlation between RDF similarity of corticosteroides and their CBG activity.

$$\begin{aligned} pK = \; & -4.64(\pm 0.24) + 2.21(\pm 0,39)H^6_{18} - 3.05(\pm 0,31)H^2_{25} \\ & (R^2 = 0.82,\ F = 64.93,\ s^2 = 0.22,\ Q^2 = 0.79); \end{aligned} \tag{6}$$

$$\begin{aligned} pK = \; & -4.63(\pm 0,20) + 2.17(\pm 0,34)H^6_{18} - 2.77(\pm 0.25)H^2_{22} \\ & -0.86(\pm 0.24)H^5_{29} \\ & (R^2 = 0.87,\ F = 55.90,\ s^2 = 0.17,\ Q^2 = 0.84); \end{aligned} \tag{7}$$

$$\begin{aligned} pK = \; & -5.51(\pm 0.12) - 1.61(\pm 0.21)H^5_{20} - 1.69(\pm 0.24)H^5_{23} \\ & +1.16(\pm 0.24)H^4_5 - 0.96(\pm 0.25)H^4_6 \\ & (R^2 = 0.91,\ F = 62.14,\ s^2 = 0.13,\ Q^2 = 0.87); \end{aligned} \tag{8}$$

$$\begin{aligned} pK = \; & -5.60(\pm 0.10) + 0.80(\pm 0.19)H^7_4 + 1.34(\pm 0.16)H^4_{18} \\ & -2.23(\pm 0.25)H^4_6 - 1.29(\pm 0.24)H^4_{24} - 0.79(\pm 0.22)H^5_{10} \\ & (R^2 = 0.93,\ F = 70.10,\ s^2 = 0.09,\ Q^2 = 0.90). \end{aligned} \tag{9}$$

Hodgkin type index H^6_{18} appearing as a descriptor in equations 6 and 7 for example represents

a measure of similarity between 6-th interval of RDF of molecule 18 and the corresponding interval of RDF of any other molecule in set. From the two parameter model (equation 6) it can be seen that the RDF code of active compound should have the dissimilar shape than the corresponing one of compound 18 in 6-th interval (8.0-9.6 Å) and similar shape than one of compound 25 in 2-nd interval (1.6-3.2 Å). Shape of RDFs on the 6-th interval is result of long distance correlation between atoms. We can expect that an important contribution to RDF shape in this interval arose from correlation between atoms that are close to positions C_3 and C_{17} (The distance between these two atom centers is about 8.5 Å for given set of steroids). Similar relative position of neighborhood atoms results in similar contribution to shape of RDFs on 2-nd interval therefore active compound must have similar decoration of steroid skeleton as compound 25 (epicorticosterone). From statistical results attached to equations from 6 to 9 it can be observed that adding of new parameters have effect to improvement of correlation between RDF similarity indices and CBG activity. On the other hand selecting a combination of few descriptors from the pool of 217 in order to find best fit for 31 propeties is risky due to chance correlation, therefore equations 6 to 9 are far less significant than they seem to be [21].

Figure 2 shows the correlation between experimental and calculated values of CBG activity for set of steroids. The values were calculated using equation 9.

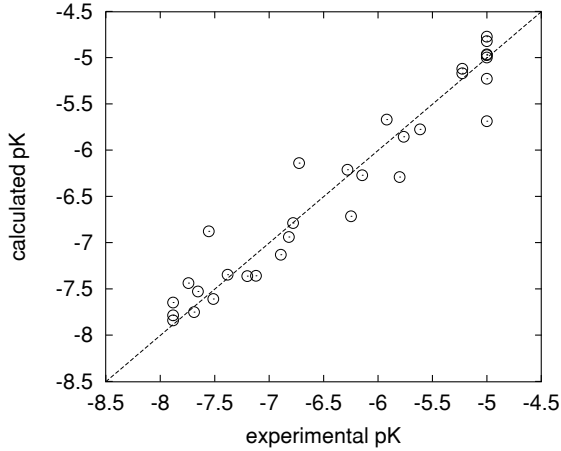


Figure 2: Correlation between experimental and calculated values of CBG activity for set of steroids. The calculated values were computed by equation 9.

4 Conclusion

The calculation of molecular similarity indices with comparison of radial distribution function is very fast and straightforward. The important feature of RDF code is rotational and translational invariant, therefore optimization of partial RDF codes alignment is not needed. The RDFs used here are based on interatomic distances thus the similarity indices from RDF are very sensitve to the change of molecular conformation. Applicability of our approach for generating QSAR models is limited due the fact that the number of generated RDF similarity descriptors is 7 fold bigger than number of compounds. In such case there is always risk of chance correlation and consequently missleading values of conventional F.

References

[1] C. Podlipnik, T. Solmajer, J. Koller, A new method for indirect evaluation of molecular shape similarity., *MATCH Commun. Math. Comput. Chem.* 52 (2004) 55–63.

[2] V. Steinhauer, 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.

[3] J. Karle, Applications of mathematics to structural chemistry, *J. Chem. Inf. Comput. Sci.* 34 (1994) 381–390.

[4] R. D. Cramer III, D. E. Patterson, J. D. Bunce, Comparative molecular field analysis (COMFA). 1. Effect of shape on binding steroids to carrier protein., *J. Am. Chem. Soc.* 110 (1988) 5959–5967.

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

[6] D. Robert, L. Amat, R. Carbo-Dorca, Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures. prediction of the corticosteroid-binding globulin affinity for a steroid family, *J. Chem. Inf. Comput. Sci.* 39 (1999) 333–344.

[7] X. Girones, R. Carbo-Dorca, Molecular quantum similarity-based QSARS for binding affinities of several steroid sets., *J. Chem. Inf. Comput. Sci.* 42 (2002) 1185–1193.

[8] H. Kubinyi, H. F. A, T. Mietzner, Three-dimensional quantitative similarity - activity relationships (3D QSIAR) from seal similarity matrices, *J. Med. Chem.* 41 (1998) 2553–2564.

[9] M. Wagener, J. Sadowski, J. Gasteiger, Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic ah receptor activity by neural networks, *J. Am. Chem. Soc.* 117 (1995) 7769–7775.

[10] http://www.chemie.uni-erlangen.de/services/steroids.

[11] J. Gasteiger, M. Marsili, Iterative partial equalization of orbital electonegativity - A rapid access to atomic charges, *Tetrahedron* 36 (1980) 3219–3228.

[12] Molecular Networks, Molecular Networks GmbH, Computerchemie, Naegelsbachstrasse 25, D-91052 Erlangen, Germany, PETRA; Fast Calculation of Physicochemical Effects in Molecules (2002).

[13] B. H. B. P. A. Kollman, K. A. Merz, Atomic charges derived from semiempirical methods, *J. Comp. Chem.* 11 (1990) 151–164.

[14] R. S. Mulliken, Electronic populations analysis on LCAO-MO molecular wave functions, *J. Chem. Phys.* 23 (1955) 1833–1840.

[15] Ampac 7.0, Semichem, Inc, Shawnee Mission, KS, 2001.

[16] M. J. S. Dewar, E. G. Zoebich, E. F. Healy, J. P. P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model, *J. Am. Chem. Soc.* 107 (1985) 3902–3909.

[17] R. Carbo, L. Lleyda, 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] http://openbabel.sourceforge.net

[20] Codessa 2.6, Semichem, Inc, Shawnee Mission, KS, 2001.

[21] D. J. Livingstone, D. W. Salt, Judging the significance of multiple linear regression models, *J. Med. Chem.* 48 (2005) 661–663.