October 2007 Volume 2 Issue 3

Physical CHEMISTRY
An Indian Journal

Trade Science Inc.

Full Paper

PCAIJ, 2(3), 2007 [171-175]

Smiles Based Optimal Descriptors: QSAR Modeling Mutagenicity

E.A.Castro[1]*, A.P.Toropova[2], A.A.Toropov[2], S.Kh.Maksudov[2]

[1]INIFTA, Chemistry Department, Faculty of Exact Sciences, La Plata National University, Diag-113 and 64,
Suc.4, C.C. 16, La Plata 1900, Buenos Aires, (ARGENTINA)

[2]Republic Uzbekistan Academy of Sciences Geology and Geophysics Institute, 700049,
Abdullaev Street 41, Tashkent, (UZBEKISTAN)

E-mail : castro@quimica.unlp.edu.ar ; eacast@gmail.com

Received: 9th June, 2007 ; Accepted: 14th June, 2007

ABSTRACT

The QSAR (Quantitative structure-activity relationship) analysis of mutagenicity of 16 dental monomers has been carried out by means of optimal descriptors calculated with SMILES (Simplified molecular input line entry system) notation. Statistical characteristics of these are n=11, r^2=0.67, s=0.59, F=18(training set); n=5, r^2=0.87, s= 0.46, F=20(test set).

KEYWORDS

QSAR;
SMILES Code;
Mutagenicity;
Salmonella test TA100.

INTRODUCTION

Mutagenicity is defined as the degree or measure of the ability to cause mutation, or, alternatively as the capacity of a physical or chemical agent to cause permanent genetic alterations. A more detailed definition of mutagenicity states that a pure substance or tested mixture is a poisonous and infectious material if there is epidemiological evidence that shows a causal connection between exposure of persons to the substance or mixture and heritable genetic effects; or there is evidence of mutagenicity in mammalian gem cells *in vivo* as shown by positive results in a study that measures mutations transmitted to offspring, or positive results in an *in vivo* study showing chemical interaction with the genetic materials of mammalian cells and positive results in an *in vivo* study assessing either gene mutation or chromosomal aberration in somatic cells.

The evidence referred to in paragraph b shall be obtained in accordance with test methods described in the ‘Introduction to the OECD guidelines on genetic toxicology testing and guidance on the selection and application of assays’, dated March 1, 1987, published in the third addendum to the OECD guidelines for gesting of chemicals; and using testing strategies described in the guidelines on the use of mutagenicity tests in the toxicology evaluation of chemicals, dated 1986, published under the authority of the minister of national health and welfare and the minister of the environment[1].

Then, understanding and predicting the chronic toxic effects of chemicals, especially mutagenicity and carcinogenicity has become one of the major problems faced by chemists involved in the development of industrial chemicals, as well as by scientists studying the toxicity of natural and xenobiotic products. The activity of a chemical towards living organisms depends upon the physical or chemical action on biological tissues, and the nature of such action will depend ultimately on the molecular structure of that chemical. This was recognized more than 100 years ago, and since then and in special in the last two decades, many attempts have been made to relate the biological activity to molecular structure in a suitable (i.e. qualitative or/and quantitative) way[2].

In order to assess the importance of this issue it is worth mentioning that, for example, in Japan based on the industrial safety and health law(ISHL) amended in 1979, manufacturers and importers, who introduce any new work place chemicals, are required to conduct bacterial mutagenicity tests and to submit the reports to the ministry of labour of Japan. By the beginning of the year 2001, the number of tested for mutagenicity and registered new chemicals had exceeded ten thousand[3]. It is also important to know that the US food and drug administration(FDA), center for drug evakluatino and research(CDER),office of pharmaceutical sciences (OPS), Informatics and computational safety analysis staff(ICSAS) is an applied regulatory research unit that compiles toxicology and safety related databases as a toxicological resource for the agency. ICSAS also produces databases suitable for quantitative structure activity(QSAR) modeling and uses these transformed databases to develop toxicology prediction software and to evaluate commercial QSAR, SAR, and data mining software to meet the needs of the FDA, other regulatory agencies, and the scientific community. These efforts are accomplished through research collaborations with software developers leveraging arrangements such as material transfer agreements(MTAs) and Cooperative research and development agreements (CRADAs). ICSAS' mission is to develop a complete battery of predictive software for all major toxicology studies recommended by the FDA's Centers. The software can be used to: (a) Improve mead compound selection by identifying and eliminating compounds with potentially significant adverse properties early in the drug discovery and development process; (b) Reduce the use of animals in testing by eliminating non-critical laboratory studies; (c) Facilitate and accelerate the review process by making better use of accumulated scientific knowledge (regulatory decision support); and (d) expand the role of QSAR and predictive toxicology by encouraging the development of complementary predictive software through collaboration with software developers and the scientific community[1].

The ames *salmonella* mutagenicity assay is a short-term bacterial reverse mutation assay which was designed to detect potential mutagens[4]. It has been used as a standard tool to detect chemical mutagens ever since it was first proposed by Ames in 1975, and is now one of the most widely used *in vitro* short-term assays throughout the world[5]. The strains employed in the Ames assay are based upon several mutants of *Salmonella typhimurium* strain LT-2. The mutations used for the test have high frequencies of chemically induced reversion and low rates of spontaneous reversion. The most frequently used test strains are TA100, TA1535, TA1537, TA97 and TA98. Among them, strains TA100 and TA1535 are employed to detect those mutagens that cause base-pair substitution mutations (e.g. induce the substitution of a leucine [GAG/CTC] by a proline[GGG/CCC] on the DNA sequences[6].

The aim of the present study is to analyze the particular mutagenicity parameter defined as the slopes of revertants vs. nanomoles of test chemical in the *Salmonella* test strain TA100 with the natural logarithm of the slopes ln(TA100) taken from Ref.[7], resorting to a particular QSAR model. We also want to estimate the predictive ability of SMILES based optimal descriptors in QSAR modeling of mutagenicity.

METHOD

The earliest expression of a quantitative structure-activity relationships between activity and chemical structure was published by Crum Brown and Frazer in 1868-9[8].

$$\Phi = f(C) \qquad (1)$$

where Φ is an expression of biological response and C is a measure of the 'constitution' of a compound. Perhaps the most famous examples of early QSAR are seen in the linear relationships between the narcotic action of organic compounds and their oil/water partition coefficients[9,10]. And the origins of modern QSAR may be traced to the work of Corwin Hansch who in the early 1960s proposed that biological 'reactions' could be treated like chemical reactions by the tech-

TABLE 1: Statistical characteristics of the models in three probes of the monte carlo optimization

	C_0	C_1	Nt	r^2	s	F	Nv	r^2	s	F
1	-415.515	410.645	11	0.6706	0.5899	18	5	0.8694	0.4602	20
2	-237.651	229.455	11	0.6680	0.5922	18	5	0.8609	0.5282	19
3	-206.772	199.596	11	0.6706	0.5899	18	5	0.8492	0.4823	17

TABLE 2 : Correlation weights of SMILES fragments over three probes of the monte carlo optimization

SF_k	CW in probe 1	CW in probe 2	CW in probe 3	Number of SF_k in the training set
]	1.0036031	1.0047425	1.0015136	2
[	1.0034938	1.0063708	1.0050731	2
O	0.9995247	0.9999423	0.9990024	21
H	0.9987730	1.0029147	1.0036978	2
C	1.0000059	0.9992630	1.0000404	71
@	1.0003216	0.9987183	1.0028250	4
C=C	1.0004423	1.0014567	1.0037321	24
=	0.9988146	0.9995566	1.0004248	14
3	1.0000271	0.9993011	0.9957878	6
2	1.0041110	1.0061907	1.0054154	20
1	1.0024948	1.0119682	1.0094488	22
(	0.9996765	1.0000266	0.9993291	26

niques of physical organic chemistry[11].

There are two key choices one must make when applying mathematical relationships like (1). One of them is to choose the set of independent variables denoted by C and the other one is to define the functional form of f. In the first case there are too many possibilities and it has lead to the so-called nightmare of the molecular descriptors.

Among such large number of possible options there is one particularly interesting, useful and convenient. In fact, one can resort to the concept of variable (or flexible) descriptors for C. This choice is interesting since C is determined for each calculation instead of being fixed under this approximation. It is also useful, since the application is rather simple. And in addition it is really convenient since final results are improved with respect to the application to the classical concept of fixed independent variables. Our research group have employed several times this sort of molecular descriptors and predictions have shown to be quite accurate[12-15].

We have also tested several functional forms for f (i.e. linear, quadratic, etc.), and although final results are not exactly the same, they do not change significantly, so that we report here the simplest one, i.e. linear relationship, for brevity reasons.

We resort here to the employment of a particular sort of flexible descriptor, the so-called optimal descriptors, which have been calculated as

$$DCW = \prod_{k=1}^{n} CW(SFk) \qquad (2)$$

where SF_k is the k-th SMILES fragment[16,17]. SMILES fragments have been detected from the SMILES line according to the following rules: first, system recognizes (if any) four character SMILES fragment, e.g., [N+], [O-], etc.; second, system recognizes (if any) three character SMILES fragment, e.g., C=C, C#C, etc.; third, system recognizes(if any) two characters fragment, e.g., Cl, Br, etc.; if there is not recognition of above fragment, system detects one-character fragment. $CW(SF_k)$ is correlation weights of the SF_k; n is number of SF_k in a given SMILES.

Correlation weights are calculated by monte carlo optimization procedure, correlation coefficient between the DCW and ln(TA100) over the training set is determined using the in role of target function. After having numerical values of the correlation weights one can calculate the DCW values on each structure of the training and test sets. Generalized model ln(TA100)=C_0+C_1 obtained by least squares method with structures of the training set should be validated with an external test set. Sixteen compounds under consideration have been split in training(n=11) and test(n=5) sets randomly, but in such a manner that all SF_k's take place in the training set.

RESULTS

Statistical characteristics of SMILES based one variable models over three probes of Monte Carlo optimization are shown in TABLE 1. Numerical values of the correlation weights as well as the list of the SF_k's on compounds under consideration are presented in TABLE 2. We display in TABLES 3 the DCW calculation for a SMILES.

Experimental and calculated values of the mutagenicity are shown in TABLE 4. Finally, in TABLE 5 we present the molecules chosen in this study together with their corresponding SMILES code.

The resulting mathematical model derived from our

TABLE 3 : DCW calculation for SMILES=‘COC1=CC=C (OCC2CO2)C=C1’ DCW=1.0109080

No.	SF_k	$CW(SF_k)$
1	C	1.0000059
2	O	0.9995247
3	C	1.0000059
4	1	1.0024948
5	=	0.9988146
6	C	1.0000059
7	C=C	1.0004423
8	(	0.9996765
9	O	0.9995247
10	C	1.0000059
11	C	1.0000059
12	2	1.0041110
13	C	1.0000059
14	O	0.9995247
15	2	1.0041110
16	(	0.9996765
17	C=C	1.0004423
18	1	1.0024948

numerical calculation is the following

$$\ln(TA100) = -415.515 + 410.645\ DCW \quad (3)$$

and characteristic statistical parameters for the training and test sets are, n= 1, r^2=0.6706, s=0.5899, F=18(Training set), n=5, r^2=0.8694, r^2_{pred}=0.8545, s=0.4602, F=20(Test set).

CONCLUSIONS

We have shown that SMILES based optimal descriptors obtained with eleven compounds give reasonable good prediction of the mutagenicity parameter for five external compounds. It suggests that definite structure-activity relationships exist for mutagenic compounds. Then, SMILES invariants can be useful for developing QSAR models for mutagenic toxicity in different cases.

TABLE 4 : Experimental and calculated with Eq. (2) values of the ln(TA100) for thetraining and test sets

Smiles	DCW	Exper.	Calc.	Exper.-calc.
Training set				
CC1=CC=C(OCC2CO2)C=C1	1.01139	0.86000	-0.19317	1.05317
CC(C1=CC=C(OCC2CO2)C=C1)(C)C	1.01010	-0.36200	-0.72299	0.36099
COC1=CC=C(CC2CO2)C=C1OC	1.01091	-0.93000	-0.38812	-0.54188
C1(CC2=CC=C(CC3=CC=CC=C3)C=C2)CO1	1.01098	-1.08000	-0.36233	-0.71767
C1(C3=CC=CC=C3)=CC=CC=C1CC2CO2	1.01163	0.62000	-0.09605	0.71605
OC[C@@H]1CO1	1.01061	-0.51400	-0.51466	0.00066
C1(CC2=CC=CC=C2)CO1	1.01186	-0.53600	0.00187	-0.53787
COC1=CC(CC2CO2)=CC=C1O	1.00928	-1.06000	-1.06108	0.00108
COC1=CC=C(CC2CO2)C=C1	1.01139	-0.89600	-0.19317	-0.70283
C1(COC2=CC=CC=C2)CO1	1.01138	0.17200	-0.19562	0.36762
[C@@H]3(CO3)COC2=C1C=CC=CC1=CC=C2	1.01729	2.23000	2.22933	0.00067
Test set				
COC1=CC=C(OCC2CO2)C=C1	1.01091	0.11500	-0.39057	0.50557
COC1=CC=CC=C1CC2CO2	1.01204	-0.57600	0.07567	-0.65167
CC1=CC=C(CC2CO2)C=C1	1.01187	-0.11100	0.00433	-0.11533
OC[C@H]1CO1	1.01028	-1.04000	-0.64808	-0.39192
[C@H]3(CO3)COC2=C1C=CC=CC1=CC=C2	1.01696	2.10000	2.09502	0.00498

TABLE 5 : Names of compounds under consideration and their SMILES code

Chemical name	SMILES
4-Methoxyphenyl glycidyl ether	COC1=CC=C(OCC2CO2)C=C1
4-Methylphenyl glycidyl ether	CC1=CC=C(OCC2CO2)C=C1
4-t-Butylphenyl glycidyl ether	CC(C1=CC=C(OCC2CO2)C=C1)(C)C
m,p-Dimethoxyphenyl propylene oxide	COC1=CC=C(CC2CO2)C=C1OC
o-Methoxyphenylpropylene oxide	COC1=CC=CC=C1CC2CO2
p-Benzylphenylpropylene oxide	C1(CC2=CC=C(CC3=CC=CC=C3)C=C2)CO1
p-Biphenylpropylene oxide	C1(C3=CC=CC=C3)=CC=CC=C1CC2CO2
R-Glycidyl alcohol	OC[C@@H]1CO1
Phenylpropylene oxide	C1(CC2=CC=CC=C2)CO1
p-Hydroxy-m-methoxyphenyl propylene oxide	COC1=CC(CC2CO2)=CC=C1O
p-Methoxyphenylpropylene oxide	COC1=CC=C(CC2CO2)C=C1
p-Methylphenylpropylene oxide	CC1=CC=C(CC2CO2)C=C1
Phenoxypropylene oxide	C1(COC2=CC=CC=C2)CO1
R-Naphthyl glycidyl ether	[C@@H]3(CO3)COC2=C1C=CC=CC1=CC=C2
S-Glycidyl alcohol	OC[C@H]1CO1
S-Naphthyl glycidyl ether	[C@H]3(CO3)COC2=C1C=CC=CC1=CC=C2

We find that resorting to rather elaborate functional forms for function f do not improve significantly final results with respect to the simple linear equation.

REFERENCES

[1] http://www.hc-sc.gc.ca/ewh-semt/pubs/occup-travail/whmis-simdut/ref_man/cpr-rpc.

[2] C.Hansch, A.Leo; ACS Professional Reference Book, American Chemical Society, Washington, DC **(1995)**.

[3] K.Sawatari, Y.Nakanishi, T.Matsushima; Ind.Health, **39**, 341-345 **(2001)**.

[4] K.Mortelmans, E.Zeiger; Mutat.Res., **455**, 29-60 **(2001)**.

[5] B.N.ames, J.McCann, E.Yamasaki; Mutat.Res., **31**, 347-364 **(1975)**.

[6] W.Barnes, E.Tuley, E.Eisenstadt; Environ.Mutagen., **4**, 297-305 **(1982)**.

[7] M.Perez-Gonzalez, A.Morales Helguera, R.Molina Ruiz, J.Garcia Fardales; Aromatic epoxides, Polymer, **45**, 2773–2779 **(2004)**.

[8] A.Crum Brown, T.Frazer; Trans.Royal Soc. Edinburgh, **25**, 151-203 **(1868-1869)**.

[9] H.Meyer; Archives of Experimental Pathology and Pharmakology, **42**, 109 **(1899)**.

[10] E.Overton; Vierteljahrsschr.Naturforsch.Ges. Zurich, **4**4, 606-15 **(1899)**.

[11] C.Hansch, R.M.Muir, T.Fujita, P.P.Maloney, F.Geiger, M.Streich; J.Am.Chem.Soc., **85**, 2817-2824 **(1963)**.

[12] E.A.Castro, A.A.Toropov, A.P.Toropova, D.V. Mukhamedzhanove; J.Arg.Chem.Soc., **93(4-6)**, 109-121 **(2005)**.

[13] P.R.Duchowicz, E.A.Castro, A.A.Toropov, E. Benfenati; S.Gupta Editor, Springer-Verlag, Heidelberg, Germany, **3**, 1-38 **(2006)**.

[14] D.J.G.Marino, E.A.Castro, A.A.Toropov; Cent.Eur. J.Chem., **4(1)**, 135-148 **(2006)**

[15] E.A.Castro, A.P.Toropova, A.A.Toropov, D.V. Mukhamedjanova; Struct.Chem., **16(3)**, 311-330 **(2005)**.

[16] D.Vidal, M.Thormann, M.Pons; J.Chem.Inf.Model, **45**, 386-393 **(2005)**.

[17] A.A.Toropov, A.P.Toropova, D.V.Mukhamedzhanova, I.Gutman; Indian J.Chem., **44A**, 1545-1552 **(2005)**.