J Mol Model (2003) 9:242–247
DOI 10.1007/s00894-003-0136-y

ORIGINAL PAPER

Mohamed Zahouily · Ahmed Rayadh · Mina Aadil · Driss Zakarya

Quantitative structure–diastereoselectivity relationships for arylsulfoxide derivatives in radical chemistry

Received: 21 October 2002 / Accepted: 3 April 2003 / Published online: 24 May 2003

Abstract Quantitative structure–diastereoselectivity relationships were studied for the intermolecular radical addition of deuterium and allyltributyltin to chiral arylsulfoxides by means of multiple linear regression and artificial neural networks (ANN). The values of diastereoselectivity (*%syn*) of the compounds studied were well correlated with the descriptors encoding the chemical structure. Using the pertinent descriptors revealed by the regression analysis, a square correlation coefficient of 0.9577 (s=5.3825) for the training set was obtained for the ANN model in a 2–4–1 configuration. The results obtained from this study indicate that the diastereoselectivity of arylsulfoxide derivatives is strongly dependent on the shape of the R and X groups.
Figure General structure of α-sulfinyl radicals

Keywords Structure–diastereoselectivity · Arylsulfoxides · Multiple linear regression · Artificial neural network · Descriptors contribution

Introduction

The control of stereoselectivity in the intermolecular reactions of acyclic radicals is an interesting field of research [1, 2, 3]. The use of sulfoxides to induce stereoselectivity for radical reactions has recently attracted much attention [4, 5, 6, 7, 8]. The stereoselectivity depends essentially on the geometric properties and physicochemical characters of the substituents attached to the α-sulfinyl radicals.

Due to the development of correlation analysis in organic chemistry, the establishment of structure–chemical behavior relationships has become a very interesting field that has lead to efficient organic synthesis [9, 10]. Indeed, the model obtained may be used as an aid for new synthetic routes or the understanding of reaction mechanisms [11, 12].

In the present work, a combination of multiple linear regression (MLR) [13] and artificial neural network (ANN) [14, 15] techniques was used for modeling the observed diastereoselectivity of the reaction of arylsulfoxide radicals [16, 17, 18, 19] with allyltributyltin and deuteriumtributyltin (Table 1). The *%syn* was considered as a diastereoselectivity index.

As a consequence, there are two main effects that should be revealed by this study:

1. The influence of the substituent effect at sites R, X and Z
2. The effect of interaction for two distinct substituents (R and X)

M. Zahouily (✉) · A. Rayadh · M. Aadil · D. Zakarya
Laboratoire de Catalyse, Synthèse et Environnement,
Département de Chimie, Faculté des Sciences et Techniques,
UFR Chimie Appliquée,
B.P. 146, 20650 Mohammedia, Morocco
e-mail: zahouily@voila.fr
Fax: +212(23)315353

Material and methods

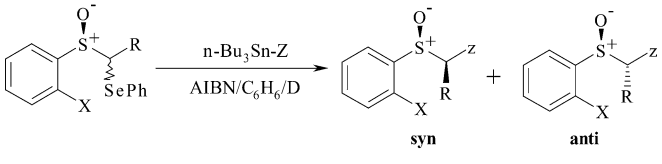
Compounds studied

The chemical structures along with the observed diastereoselectivity data of the compounds used in this study are shown in Table 1. The diastereoselectivity data were taken from various studies [16, 17, 18, 19].

Descriptors

The main step in our study was the choice of chemical structure. It is obvious that the performance of the models depends mostly on the parameters used to describe the molecular structures.

In this study, a set of descriptors related to physicochemical and geometric properties of the molecules was used. All these



Scheme 1 Reaction of arylsulfoxides radical with allyltributyltin and deuteriumtributyltin

Table 1 Chemical structure of chiral arylsulfoxides derivatives and observed diastereoselectivity (*%syn*) (see Scheme 1)

N°	R	X	Z	*%syn*[a]	Ref.
1	Me	H		61	13,15
2	Me	H		66	13,15
3	Me	H	D	47	16
4	Me	H		50	13
5	Me	Cl	D	77	16
6	Me	Cl		88	13,15
7	Me	Cl		90	13,15
8	Me	Cl		80	13,15
9	Me	OH		65	13
10	Me	O-THP		79	13
11	Me	O-MOM		82	13
12	Me	O-nPr		81	13
13	Me	O-Piv		71	13
14	Me	O-TBDMS		68	13
15	Me	O-Ts		81	13
16	Me	Br		90	13
17	Et	H		39	16
18	Et	Cl	D	63	16
19	Et	Cl		72	16
20	*i*-Pr	H	D	23	16
21	*i*-Pr	H		22	16
22	*i*-Pr	Cl	D	37	16
23	*i*-Pr	Cl		40	16
24	*t*-Bu	H	D	10	16
25	*t*-Bu	H		05	16
26	CF_3	H		19	14
27	CF_3	H		18	14

[a] *%syn*: indicating the diastereomer percentage observed.

descriptors were calculated for the substituents R, X and Z (Table 1).

The descriptors given below were calculated with pro-Demo (TM) Revision 3.01 demo published by ChemSW Software (TM) [20].

- Size and shape described by means of van der Waals volume (V) and van der Waals surface (S).
- Molecular dimensions (length, width and height). Length (L) is the distance along the screen x-axis between the left and rightmost atoms plus their van der Waals radii. Width (W) is the distance along the screen y-axis between the top and bottom-most atoms plus their van der Waals radii. Height (H) is the distance along the screen z-axis between the nearest and farthest atoms plus their van der Waals radii.
- log P, the partition coefficient between n-octanol and water.
- Molar refractivity (MR).
- Molecular weight (MW).
- Hydrogen-bonding donors (HBD), hydrogen-bonding acceptors (HBA).

Some other descriptors, V/L, V/W, W/H and ovality were calculated on the basis of the descriptors elaborated.

- Ovality estimation (O), for each substituent was that given by Bodor [21].
- $O = S/(4\pi K)$ (1a)
- $K = (3V/4\pi p)^{2/3}$ (1b)

Data analyses

Multiple linear regression (MLR)

This method was used to generate linear models between the diastereoselectivity (*%syn*) and the molecular descriptors used. Because of the large number of descriptors considered, a stepwise procedure combining the forward and backward algorithms was used to select the pertinent descriptors.

In order to avoid all difficulties in the interpretation of the resulting models, pairs of variables with a correlation coefficient greater than 0.70 were considered as intercorrelated. In such a situation, only one was included in the screened model. The quality of the model was considered as statistically sufficient on the basis of the squared correlation coefficient (r^2), standard deviation (s), and F-statistics (F) when all parameters in the model were significant at 95% confidence level ($p<0.05$).

The cross-validation (CV) procedure was employed after variable selection to test the validity and predictive ability of the models.

In this work the leave-one-out procedure was used to evaluate the predictive ability of the MLR and ANN. The cross-validation coefficient q^2 was calculated according the following equation: [22]

$$q^2 = 1 - (\mathrm{PRESS}/\mathrm{Variance}) \quad (2)$$

where PRESS means predictive residuals.

Artificial neural networks (ANN)

The application of ANNs to solve problems in chemistry is a recent field of research. ANNs have been applied to the investigation of QSARs [23, 24, 25, 25].

Neural networks models are known to be very effective in representing the nonlinear relationships between variables in complex systems. Most of the applications of ANNs to chemistry use the back-propagation algorithm (BPA) [26]. Consequently, it has been employed in the present study.

Results and discussion

Multiple linear regression analysis

Multiple linear regression analysis was performed on the compounds described in Table 1, a few suitable models were obtained and the best one was selected and presented in Eq. (3):

$$\%syn = 91.5079(\pm 4.5226) - 14.2756(\pm 1.0485)S(\mathrm{R}) + 1.5875(\pm 0.2689)V/L(\mathrm{X}) \quad (3)$$
$$n = 27 \quad r^2 = 0.8486 \quad s = 7.7891$$

The statistical quality of Eq. (3) is moderate and accounts for 85% (r^2=0.8486) of the information represented by data. The high diastereoselectivity is associated with low surface [S(R)] with increased shape [V/L(X)]. The calculated contributions [27] for descriptors S(R) and V/L(X) were 70% and 30% respectively. There are two compounds with a large estimation error for Eq. (3) (compounds 3 and 14), and when these compounds were excluded the standard deviation goes from 7.7891 to 6.3051.

The plot of experimental versus calculated diastereoselectivity is given in Fig. 1a. Cross-correlation analysis showed that all pairwise correlations were $\leq$0.6681 in this equation, also indicating a low collinearity (see Table 2).

Cross-validation

In the cross-validation phase, 27 subsets were created according to the leave-one-out method and the output of the removed compound was predicted for each subset [14]. They yielded a q^2=0.898, indicating a good quality of the model according to Wold [28].

Artificial neural network (ANN)

In order to test the possibility of nonlinear effects on the data and to establish a more accurate model, we used a neural network technique [29, 30].

The ANN [20] was trained by the back-propagation (BP) of errors algorithm [14] and had the following architecture:

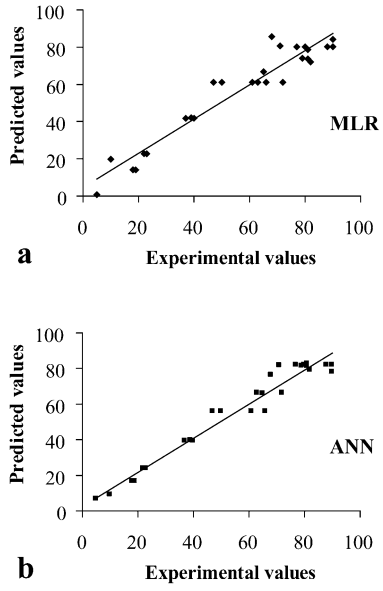


Fig. 1a, b Experimental and predicted values from MLR (**a**) and ANN (**b**) respectively for training sets

Table 2 Correlation matrix

	%syn	*S*(R)	*V/L*(X)
%syn	1	0.9409	−0.7695
S(R)		1	−0.6681
V/L(X)			1

- An input layer including pertinent descriptors from the MLR.
- A hidden layer for which the ratio of the number of data points in the training set and the number of connections controlled by the network, ρ, is critical to the predictive power of the neural net. The range $1.8<\rho<2.2$ [ρ=(number of data points in the training set)/(number of adjustable weights controlled by the network)] [15] was used as a guideline for an acceptable number of neurons in the hidden layer. It is claimed that for $r \ll 1.0$ the network simply memorizes the data, whereas for $r \gg 3.0$ the network loses its ability to generalize.
- Output layer of one neuron, representing the diastereoselectivity (*%syn*). The input values were normalized.

All the trials were made for 1,000 iterations and repeated ten times to make sure of the reproducibility of results. The connection weights were stable after 800 iterations.

The best model is that corresponding to the optimum r^2 and s parameters (see Table 3). A hidden layer of four neurons was selected.

We obtained a square correlation coefficient of 0.9577 (n=27) between calculated and observed diastereoselectivity (*%syn*) with a standard deviation of 5.3825.

Table 3 Variation of r^2 and s with the number of neurones of the hidden layer

Hidden neurones	r^2 (training)	s (training)
2	0.9567	5.4494
3	0.9573	5.4071
4	0.9577	5.3825
5	0.9577	5.3837
6	0.9577	5.3838
7	0.9577	5.3884

This preliminary study enables us to conclude that the ANN with (2–4–1) architecture was able to establish a satisfactory relationships between the pertinent descriptors and the diastereoselectivity of arylsulfoxide derivatives.

We used the same procedure as for the MLR analysis and obtained a coefficient of cross-validation equal to q^2=0.918. The model obtained was considered to be good predictive one, according to Wold [28]. The performances of the ANN are superior to that of MLR and this indicates the presence of nonlinearity in the data since the efficiency of descriptors was increased. The combination between MLR and ANN was fruitful.

The plot in Fig. 1b indicates that there is a significant correlation between actual values and calculated values of diastereoselectivity (*%syn*) from the ANN for the training sets.

Analysis of descriptor's contribution in ANN model

To estimate the relative contribution of descriptors, we chose two different approaches:

(i) The contribution of descriptors i (i=1–2) was estimated from the trained 2–4–1 configuration network. The descriptor under study was removed from the 1–4–1 ANN together with its corresponding weights. Then the network (1–4–1) calculated the output of each molecule as usual. The mean of the deviations of the absolute values Δm_i between the observed diastereoselectivity and the estimated diastereoselectivity for all compounds was calculated. This process was reiterated for each descriptor. Finally, the contribution C_i [31] of descriptor i is given by:

$$C_i = 100 \times \Delta m_i / \sum_{i=1}^{2} \Delta m_i \qquad (4)$$

(ii) We analyze deviations when a given descriptor is removed and for the full set of descriptors. This approach is an extension of the previous one proposed by Chastrette et al. [32]. In this way we could estimate the contribution of each descriptor removed from the model. Table 4 shows that the two methods give the same results.

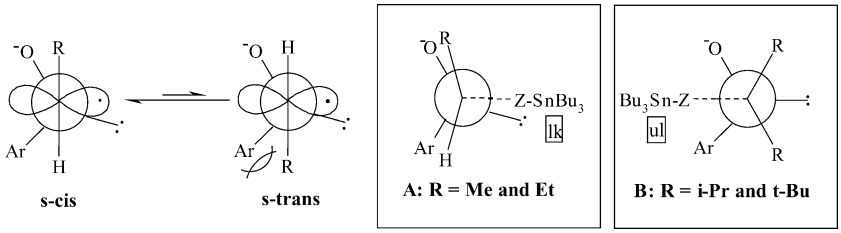


Fig. 2 The most reactive s-*cis* and s-*trans* conformations of the intermediate α-sulfinyl radicals

Table 4 Evaluating the impact of each descriptor in ANN

Removed descriptor	[a]*C%*	[b]r^2	[b]*s*
S(R)	63	0.4181	19.9580
V/L(X)	37	0.8129	11.3141
Any one	100	0.9577	5.3825

[a] The contribution (*C%*) of descriptor given by the first method described in the text
[b] Given by the second method described in the text

These results, indicate that the relative importance of the descriptors varied in the following order: *S*(R)>*V/L*(X). These calculations confirm the great effect of the substituents R and X on the diastereoselectivity. The information obtained from the established model can eventually allow us to predict the diastereoselectivity of arylsulfoxides [7, 8].

The α-sulfinyl radicals possess two minimum-energy conformations s-*cis* and s-*trans* (Fig. 2), the s-*cis* being more stable by 1.0–3.5 kcal mol^{-1}. [19] These two conformations correspond to those predicted by stereoelectronic interactions, the singly occupied orbital is perpendicular to the S–O bond for optimal overlap [33], and by the minimization of steric and eclipsing interactions.

Therefore, the repulsion steric interactions between the methyl group and a substituent X in the *ortho* position of the phenyl ring destabilized the s-*trans* conformation of the radical relative to the s-*cis*. With secondary and tertiary R groups, the s-*cis* conformation is more stable because of strong steric interactions between the R group and the aryl moiety in the s-*trans* conformer. Depending on the surface of the R groups, the s-*cis* conformer is attacked preferentially with a like (*lk*) (R=methyl, primary alkyl groups) or an unlike (*ul*) topicity (R=secondary and tertiary groups). The *lk* topicity is favored by steric factors [*S*(R) and *V/L* (X)] since the attack occurs *anti* to the bulky aryl group (transition state **A**). Due to pyramidalization in the transition state [34], the *lk* attack generates eclipsing interactions between the R group and the oxygen atom at sulfur. These eclipsing interactions become dominant with large surface alkyl groups. Attack from the more hindered face (*ul* topicity) leading to the staggered transition state **B** becomes more favorable (Fig. 2).

To ensure that the results obtained in MLR and ANN were not due to chance and to lend credence to our results, we have run a scrambling experiment [21, 22, 23]. The dependent variable *%syn* is randomly scrambled and then the same algorithms used in MLR and ANN run once again. The statistical results as the correlation coefficient square r^2 and the standard deviation *s* of its results are compared with the r^2 and *s* of the MLR and ANN models developed in this work. The r^2 values were 0.2049 and 0.6303 compared with 0.8486 and 0.9577 for the *s* values we obtained 24.7436 and 15.9076 compared with 7.7891 and 5.3825 for the training set in MLR and ANN, respectively. This test confirms and clearly shows that the descriptors selected in this study describe the diastereoselectivity studied very well.

Conclusion

Taking into account the complexity of the modeled diastereoselectivity, we were able to show with only two 2-D descriptors, that the diastereoselectivity of the arylsulfoxide derivatives was strongly controlled by the steric factors of the substituents attached to the α-sulfinyl radicals.

The pattern obtained with the ANN approach is more efficient than regression analysis, since it reveals the nonlinear effects in α-sulfinyl radicals. In addition, the approach used for the contributions and classification of descriptors in the ANN may be of help in quantitative structure–diastereoselectivity relationships interpretation.

References

1. Smadja W (1994) Synlett 1–12
2. Zahouily M (2002) Acad Sci Paris 5:655–658
3. Porter N, Giese B, Curran DP (1991) Acc Chem Res 24:296–299
4. Smadja W, Zahouily M, Journet M, Malacria M (1991) Tetrahedron Lett 32:3683–3686
5. Smadja W, Zahouily M, Malacria M (1992) Tetrahedron Lett 33:5511–5514
6. Tsai YM, Chang FC, Huang JM, Shiu CL, Kao CL, Liu JS (1997) Tetrahedron 53:4291–4308
7. Zahouily M, Journet M, Malacria M (1994) Synlett 366–369
8. Imboden C, Bourquard T, Corminboeuf O, Renaud P, Schenk K, Zahouily M (1999) Tetrahedron Lett 40: 495–498
9. Lehman PZF (1987) Quant Struct Act Relat 6:57–60
10. Zakarya D, Farhaoui L, Fkih-Tetouani S (1996) J Org Phys Chem 9:672–675
11. Zakarya D, Rayadh A, Zair T, Essaoudi A (1999) Acad Sci Paris 2:153–156
12. Zakarya D, Rayadh A, Samih M, Lakhlifi T (1994) Tetrahedron Lett 35:2345–2348
13. Livingstone DJ, Manallack DT (1993) J Med Chem 36:1295–1297

14. Rumhelart DE, Hinton CE, Williams RJ (1986) Nature 323:533–536
15. So S, Richards WG (1992) J Med Chem 35:3201–3207
16. Renaud P, Bourquard T (1994) Tetrahedron Lett 35:1707–1710
17. Renaud P, Carrupt PA, Grester M, Schenk K (1994) Tetrahedron Lett 35:1703–1706
18. Renaud P, Bourquard T (1995) Synlett 1021–1023
19. Zahouily M, Caron G, Carrupt PA, Knouzi N, Renaud P (1996) Tetrahedron Lett 37 8387–8390
20. Data pro Qnet 2000 for Windows V2 K build neutral network modeling. Vesta Services, Winnetka, Ill.
21. Bodor N, Harget A, Huang MJ (1991) J Am Chem Soc 113:9480–9483
22. Tetko IV, Villa AEP, Livingstone DJ (1996) J Chem Inf Comput Sci 36:794–803
23. Bazoui H, Zahouily M, Sebti S, Boulajaaj S, Zakarya D (2002) J Mol Model 8:1–7
24. Bazoui H, Zahouily M, Zakarya D, Sebti S, Boulajaaj S (2002) Phys Chem News 6:124–130
25. Zahouily M, Rhihil A, Bazoui H, Sebti S, Zakarya D (2002) J Mol Model 8:168–172
26. Normandin A, Grandjean BPA, Thibault J (1993) Ind Eng Chem Res 32:970–975
27. Gore WL (1952) Statistical methods for chemical experimentation. Interscience, New York, p 141
28. Wold S (1991) Quant Struct Act Relat 10:191–193
29. Cherqaoui D, Esseffar M, Zakarya D, Mesbah A, Villemin D (1998) Models Chem 135:79–91
30. Zakarya D, Chastrette M, Tollabi M, Fkih-Tétouani S (1999) Chemom Intell Lab Syst 48:35–46
31. Cherqaoui D, Esseffar M, Villemin D, Cence JM, Chastrette M, Zakarya D (1998) New J Chem 22:839–843
32. Chastrette M, Zakarya D, Peyraud JF (1994) Eur J Med Chem 29:343–348
33. Ianelli S, Musatti A, Nardelli M, Benassi R, Folli U, Taddei F (1992) J Chem Soc, Perkin Trans 2 49–57
34. Dewar MJS, Hwang JC, Kuhn DR (1991) J Am Chem Soc 113:735–741