12th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-12) 1-30 November 2008
http://www.usc.es/congresos/ecsoc/12/ECSOC12.htm & http://www.mdpi.org/ecsoc-12

[G0012]

STRUCTURE-AFFINITY MODELING OF AZO DYE ADSORPTION ON CELLULOSE FIBRE BY MLR

SIMONA FUNAR-TIMOFEI[a], LUDOVIC KURUNCZI[b], WALTER M.F. FABIAN[c], DANIELA IONESCU[b]

[a]Institute of Chemistry of the Romanian Academy, 24 Mihai Viteazul Bvd., 300223 Timisoara, Romania, e-mail: timofei@acad-icht.tm.edu.ro

[b]University of Medicine and Pharmacy "Victor Babes" Timisoara, Faculty of Pharmacy, 2 E. Murgu P-ta, 300041 Timisoara, Romania

[c]Institut für Chemie, Karl-Franzens Universität Graz, Heinrichstrasse 28, A-8010 Graz, Austria

ABSTRACT

Quantitative structure-affinity relationships were applied by multiple linear regression (MLR) analysis for a series of 21 monoazo dyes. Calculated 0D, 1D and 2D structural dye features were correlated to their affinity for cellulose. Variable selection was performed by the genetic algorithm. Good correlations with dye affinity and models with predictive power were obtained. Electrostatic interactions are favorable and hydrophobic disfavorable for dye binding on cellulose.

INTRODUCTION

Several computational methods have been employed in the study of textile adsorption on cellulose fibre [1, 2].

The classical QSAR methods rely principally on the mathematical technique of multiple linear regression (MLR). This means an easy interpretation of the results, especially when the fibre affinities of the dye molecules are related to simple and clearly defined physico-chemical parameters, but implies some risks of chance correlation. This disadvantage can be improved by the introduction of several criteria during the variable selection. The number of parameters potentially important for the dye fibre interaction can be large and this leads to the use of multivariate statistical methods, like principal component analysis, principal component regression analysis or PLS (projection in latent structures). These methods successfully handle large matrices of predictor variables, although sometimes with disadvantage of clarity as well as of physical and chemical interpretation.

This paper presents a quantitative structure-affinity relationships study for a series of azo dyes by the multiple linear regression (MLR) method. Structural dye features obtained by molecular modeling techniques were correlated to their affinity for cellulose. Variable selection was performed by the genetic algorithm and several MLR models were obtained. They give information on the dye adsorption mechanism on fibre.

METHODS AND MATERIALS

Molecular descriptors

A series of 21 dyes was considered, having as dependent variable the affinity for cellulose fibre taken from literature [3-5] (see table 1).

The molecular dye structures were built by the ChemOffice package [6] and energetically optimized by molecular mechanics calculations. The optimized structures were further used to derive structural dye descriptors. Several types of 0D, 1D and 2D descriptors were calculated by the Dragon software [7] : constitutional (e.g. MW-molecular weight, AMW-average molecular weight, Mp-mean atomic polarizability (scaled on Carbon atom), Me-mean atomic Sanderson electronegativity (scaled on Carbon atom), Ss-sum of Kier-Hall electrotopological states, nS-number of Sulfur atoms, SCBO-sum of conventional bond orders (H-depleted)), functional groups counts (like: nCp- number of terminal primary C(sp3) atoms, nHBonds-number of intramolecular H-bonds (with nitrogen, oxygen, fluorine), nThiazoles-number of Thiazoles, nSO2OH- number of sulfonic (thio-/dithio-) acids) and molecular properties (like: ALOGP-Ghose-Crippen octanol-water partition coefficient, TPSA(Tot)-topological polar surface area using nitrogen, oxygen, sulphur, phosphor polar contributions). Descriptors included in the final MLR models are presented in table 2.

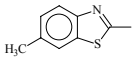
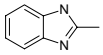
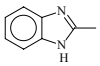
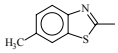
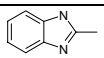
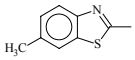
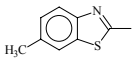
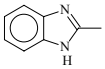
Multiple Linear Regression (MLR)

Multiple linear regression relates one experimental variable y_k to one or several structural variables x_i by the equation [8]:

$$y_k = b_o + \sum_i b_i \cdot x_{ik} + e_k \qquad (1)$$

where b represents regression coefficients and e the deviations and residuals. MLR calculations were performed by the STATISTICA package [9].

Table 1. The studied compounds and their affinities (A)

No.	Compound structure	Y	A (kJ/mole)	No.	Compound structure	Y	A (kJ/mole)
1		γ_b	22.26	11		H	9.49
2		γ_b	15.69	12		γ_b	8.58
3		γ_a	14.35	13		C	7.70
4		H	14.48	14		H	7.24
5		H	13.56	15		H	6.61
6		γ_b	13.18	16		R	5.23
7		γ_b	10.92	17		R	4.60
8		C	10.50	18		R	4.48
9		R	9.62	19		C	3.59
10		R	8.79	20		C	2.97
				21		C	1.92

[a] A - experimental affinities; Y - coupling components: γ acid coupled in acidic (γ_a), respectively basic (γ_b) medium, H - H acid, C - chromotropic acid, R - R acid

Model validation

In order to test the predictive power of the model, the following statistical measures were used [10]: 1) correlation coefficient R between the predicted and observed activities; 2) coefficient of determination for linear regressions with intercepts set to zero, i.e. R_0^2 (predicted versus observed activities), and $R_0'^2$ (observed versus predicted activities); 3) slopes k and k' of the above mentioned

two regression lines. The following conditions should be satisfied for an acceptable predictive power model:

$$q^2 > 0.5 \quad (2)$$

$$R^2 > 0.6 \quad (3)$$

$$\frac{(R^2 - R_0^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \le k \le 1.15 \quad (4)$$

or

$$\frac{(R^2 - R_0'^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \le k' \le 1.15 \quad (5)$$

$$\left|R_0^2 - R_0'^2\right| < 0.3 \quad (6)$$

In addition to these criteria, Q^2_{ext} values were calculated by the MobyDigs software [11] to test the predictive power of the model obtained from the training set compounds. The external validation technique uses a test set to perform a further check on the predictive capabilities of a model obtained from a training set and with predictive power optimized by an evaluation set. By using the selected model the values of the response for the test objects are calculated and the quality of these predictions is defined in terms of Q^2_{ext}, which is defined as:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{ext}} \left(y_i - \bar{y}\right)^2} \quad (7)$$

where the sum runs over the test set objects (n_{ext}) and y, $\hat{y}$ and $\bar{y}$ are the experimental, predicted, respectively the average values of the training set responses.

RESULTS AND DISCUSSIONS

The series of 21 dyes was studied by molecular mechanics calculations and the optimized structures thus derived were used to calculate dye descriptors. The descriptors used in the final MLR model are presented in table 2. MLR calculations have been performed by the STATISTICA software [9].

From the entire set of dyes, a training set of 16 compounds and a test set of 5 compounds: no. 2, 6, 15, 17, 19 (see table 1) were considered. The test set compounds were selected consulting the scores scatter plots of the first three principal components (82.1 % of the variance explained) for the principal component analysis (PCA) model constructed using the matrix of the whole set of descriptor variables for the 21 analyzed compounds. We have included in the test set one of two similar compounds (grouped together) positioned on the opposite sides of the plot origin in the four

quadrants of the respective plots. PCA analysis was performed by the SIMCA-P+ software [12]. Variable selection was carried out by the genetic algorithm included in the MobyDigs program [11], using the RQK function [13], as fitness function. Leave-one-out crossvalidation and bootstrapping techniques were used for the internal validation of the obtained MLR models.

Table 2. Calculated dye descriptors: average molecular weight (AMW), number of terminal primary C(sp3) atoms (nCp), topological polar surface area using nitrogen, oxygen, sulphur, phosphor polar contributions (TPSA(Tot)), number of sulfonic (thio-/dithio-) acids (nSO2OH), number of intramolecular H-bonds (with N,O,F) (nHBonds),Ghose-Crippen octanol-water partition coefficient (AlogP)

No.	AMW	nCp	TPSA(tot)	nSO2OH	nHBonds	AlogP
1	9.43	1	174.85	1	1	5.552
2	9.19	0	162.4	1	1	4.498
3	9.19	0	162.4	1	2	4.498
4	10.29	1	211.58	2	2	5.262
5	10.09	0	199.13	2	2	4.208
6	9.31	0	162.4	1	1	2.87
7	9.55	0	175.29	1	1	2.346
8	10.19	1	237.6	2	2	4.516
9	10.39	1	231.81	2	2	4.995
10	10.2	0	219.36	2	2	3.94
11	10.32	0	199.13	2	1	2.58
12	9.8	0	188.18	1	1	2.436
13	9.99	0	225.15	2	2	3.461
14	10.57	0	212.02	2	1	2.452
15	10.83	0	224.91	2	2	2.146
16	10.44	0	219.36	2	2	2.312
17	10.94	0	245.14	2	2	1.878
18	10.68	0	232.25	2	2	1.789
19	10.2	0	225.15	2	2	1.833
20	10.44	0	238.04	2	2	1.31
21	10.68	0	250.93	2	2	1.399

Two MLR models were found to be predictive. They are presented in Table 3. Best correlations with dye affinity and statistical results were noticed in model 1.

The predictive power of the best MLR model was then checked by the criteria stated by A. Tropsha et al [10] (see equations (2) to (6)). All these calculated criteria indicated a model with predictive power, respectively:

$$Q^2_{ext} = 0.929 > 0.5$$

$$R^2 = 0.951 > 0.6$$

$$\frac{(R^2 - R_0^2)}{R^2} = 0.024 < 0.1 \quad \text{and} \quad 0.85 \le k = 1.003 \le 1.15$$

$$\frac{(R^2 - R_0'^2)}{R^2} = 0.005 < 0.1 \quad \text{and} \quad 0.85 \le k' = 0.981 \le 1.15$$

$$\text{and } \left|R_0^2 - R_0'^2\right| = 0.018 < 0.3$$

Table 3. Final MLR models for the series of 21 dyes*

No	Model	R^2	Q^2	Q^2_{boot}	Q^2_{ext}	$a(r^2)$	$a(q^2)$	F	s
1	A = -151.48 (±35.23) – 7.05 (±0.96) nSO2OH + 350.16 (±51.37) nHBonds – 0.62 (±0.23) ALOGP	0.951	0.828	0.803	0.923	0.257	-0.429	37.77	1.75
2	A = 38.96 (±4.10) + 6.85 (±1.12) nCp - 0.15 (±0.02) TPSA(tot)	0.874	0.809	0.822	0.897	0.056	-0.403	45.02	1.93

* R^2 - squared multiple regression coefficient, Q^2 - leave-one-out cross-validated R^2, Q^2_{boot} - bootstrapping Q^2, Q^2_{ext} - external Q^2 (for the test set), Y-scrambling parameters [14] (a(r2), a(q2)), F- Fischer test, s- standard deviation

Hydrogen bonds between dye and cellulose are expected to have highest contribution to the dye affinity. Dye sulfonic acid groups and dye hydrophobicity are detrimental for the dye binding. Dye polar surface area decrease the dye affinity, being probably related to the hydrophobic interactions at the dye surface-dyebath solution interface.

CONCLUSIONS

Dye binding to cellulose was studied by correlations of dye affinity values with dye descriptors by the multiple linear regression (MLR) method. Dye structures were modeled by molecular mechanics and 0D, 1D and 2D descriptors were derived from the optimized structures.

The dye affinity for cellulose increases with the increased number of hydrogen bonds. Dye sulfonic acid groups, hydrophobicity and polar surface area disfavored the dye binding to cellulose. Sulfonic acid groups probably participate only to dye solubilizing in the dyeing environment.

ACKNOWLEDGEMENTS

This project was financially supported by Ministerul Educatiei, Cercetarii si Tineretului - Autoritatea Nationala pentru Cercetare Stiintifica (MEdC-ANCS), contract grant number: 71GR/2006.

REFERENCES

1. Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z. Dyes Pigm 2000; 47: 5-16.
2. Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A.; J Chem Inf Comput Sci 2004; 44: 1423-1435.
3. Alberti, G.; Cerniani, A.; De Giorgi, M.R.; Seu, G. Red Sem Fac Sci 1978; 48: 267-273.
4. Alberti, G.; Cerniani, A.; De Giorgi, M.R.; Seu, G. Ann Chim (Rome) 1981; 295-298.
5. Alberti, G.; Cerniani, A.; De Giorgi, M.R.; Seu, G. Teintex 1981; (1-2): 17-26.
6. Chem3D Ultra 6.0, CambridgeSoft.Com, Cambridge, MA, U.S.A.
7. Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy
8. Wold, S.; Dunn III, WJ. J Chem Inf Comput Sci 1983; 23: 6-13.
9. STATISTICA 7.1, StatSoft Inc, Tulsa, OK, USA
10. Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.-D.; Lee, K.-H.; Tropsha, A. J Comput Aid Mol Des 2003; 17: 241-253.
11. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. MobyDigs: Software for Regression and Classification Models by Genetic Algorithms, in: 'Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks'. (R. Leardi Ed.), Chapter 5, Elsevier, 2003, p. 141-167.
12. SIMCA-P+, 12.0.0.0, 2008, Umetrics A.B., Sweden
13. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Anal Chim Acta 2004; 515: 199-208.
14. Lindgren, F.; Hansen, B.; Karcher, W.; Sjöström, M.; Eriksson L. J Chemometr 1996; 10: 521-532.