13rd International Electronic Conference on Synthetic Organic Chemistry (ECSOC-13), 1-30 November 2009.
http://www.mdpi.org/ecsoc-13/ & http://www.usc.es/congresos/ecsoc/13/

[G006] COMPARISON OF SEVERAL REGRESSION METHODS APPLIED IN DISPERSE DYE-CELLULOSE BINDING

SIMONA FUNAR-TIMOFEI

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

ABSTRACT

Quantitative structure-affinity relationships were applied to a series of 27 disperse dyes by partial least squares (PLS) analysis and compared to previously published MLR (multiple linear regression), MTD (minimum steric difference) and CoMFA (comparative molecular field analysis) results. Calculated 0D, 1D and 2D structural dye features were correlated to their affinity for cellulose by PLS. A robust model ($R^2X(cum) = 0.617$, $R^2Y(cum) = 0.959$, $Q^2(cum) = 0.953$) with predictive power was obtained from these correlations. Better statistical results were achieved in the PLS model, in comparison to the previous MLR, MTD and CoMFA results, but the three-dimensional models obtained by CoMFA gave more information on the dye-cellulose specific interactions.

INTRODUCTION

Disperse dye adsorption was studied mostly for cellulose acetate and triacetate, nylon, polyethylene terephthalate and acrylic fibres, but it was found that these dyes can, also, be adsorbed by cellulose to some extent [1]. In case of cellulose dyeing by some 4-aminoazobenzene dyes, it was found that there was no evidence of hydrogen bonding between these dyes and the fibre; the attraction forces could be explained by the dipole forces on a region of cellulose where water molecules are absent [2].

Previous QSAR studies of disperse dye structure-affinity to cellulose fibre were reported [3-5]. Several methods were applied to a series of 20 dyes to quantify structure-affinity relationships, like: Free-Wilson and MTD (minimum steric difference) [3]. MLR (multiple linear regression) approach was applied to model the dye-cellulose binding by correlations of dye affinity to several parameters, like: the sum of π-Hansch substituent term ($\sum\pi$), the sum of Hammett substituent constants ($\sum\sigma$), sum of molar refractivities, of Charton steric substituent constant and Verloops

Sterimol parameters [4]. A satisfactory MLR (multiple linear regression) equation was obtained (r = 0.854, s = 1.03, q^2_{LOO} (leave-one-out crossvalidation coefficient) = 0.590) for 19 disperse dyes, indicating the influence of hydrophobic and electronic interactions in dye-fibre binding.

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 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.

In this paper results obtained by PLS are compared to previous MLR, MTD and CoMFA published results obtained for the adsorption on cellulose of 27 disperse dyes. 0D, 1D and 2D structural dye features obtained by molecular modeling techniques were correlated to their affinity for cellulose by PLS.

METHODS AND MATERIALS

Molecular descriptors

A series of 27 dyes was considered, having as dependent variable the affinity (table 1) for cellulose fibre taken from literature [2, 6].

The molecular dye structures were built by the ChemOffice package [Chem3D Ultra 6.0, CambridgeSoft.Com, Cambridge, MA, U.S.A.] and energetically optimized by molecular mechanics calculations. The optimized structures were further used to derive structural dye descriptors. 76 descriptors were calculated by the Dragon software [Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy]: constitutional, functional groups counts and molecular properties (of 0D, 1D and 2D type).

The Partial Least Squares (PLS) method

Projections to Latent Structures (PLS) represent a regression technique for modeling the relationship between projections of dependent factors and independent responses. PLS (Partial Least Squares) regression is a statistical modeling technique with data analysis features linking a block (or a column) of response variables to a block of explanatory variables [7]. The PLS approach leads to stable, correct and highly predictive models even for correlated descriptors [8].

This method describes the matrix X, of chemical descriptors of the training set (N compounds) defining a number of F significant principal components (PC), i.e. t_{if} columns formed by equation (1), when i = 1, ..., N.

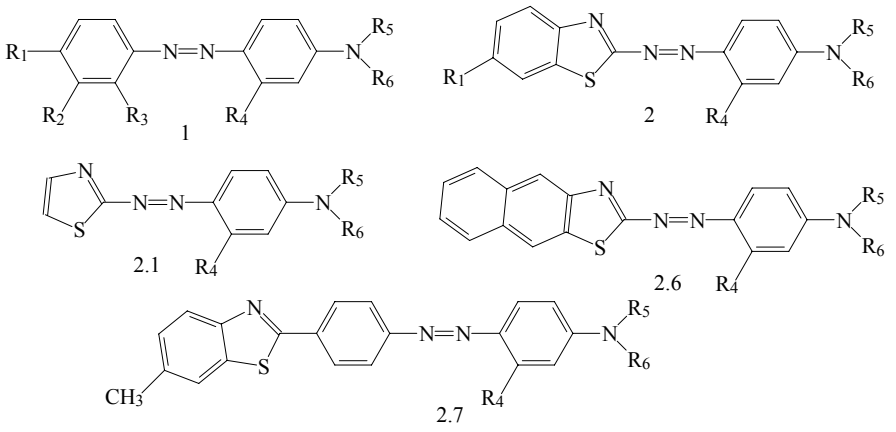
$$x_{ik} = \bar{x}_k + \sum_{f=1}^{F} p_{fk} \cdot t_{if} + e_{ik} \quad (1)$$

where $\bar{x}_k$ denotes the mean of variable k, p_{fk} the loading of variable k in dimension (factor) f, and e_{ik} the residuals [9]. The consecutive orthogonal latent variables (t_f) are deduced assuring maximal covariance of these with y. The linear PLS inner relation is described by equation (2):

$$y_i = \bar{y} + \sum_{f=1}^{F} b_f \cdot t_{if} + e_i \quad (2)$$

where $\bar{y}$ represents the average of the y-variable and b_f the regression coefficients. These can be transformed to express the biological activity y in function of the original x_k descriptors.

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



Cpd.	R_1	R_2	R_3	R_4	R_5	R_6	A (kJ/mole)	Cpd.	R_1	R_2	R_3	R_4	R_5	R_6	A (kJ/mole)
I.1	NO_2	H	H	H	C_2H_5	C_2H_4OH	11.06	I.15	CH_3	H	H	H	C_2H_4OH	C_2H_4OH	5.93
I.2	NO_2	H	H	CH_3	C_2H_4OH	C_2H_4OH	10.89	I.16	F	H	H	H	C_2H_4OH	C_2H_4OH	5.69
I.3	NO_2	H	H	H	H	H	9.73	I.17	H	H	CH_3	H	C_2H_4OH	C_2H_4OH	5.69
I.4	Br	H	H	H	C_2H_4OH	C_2H_4OH	9.32	I.18	H	H	H	H	H	H	5.29
I.5	NO_2	H	H	H	C_2H_4OH	C_2H_4OH	8.15	I.19	H	H	H	H	C_2H_4OH	C_2H_4OH	4.61
I.6	Cl	H	H	H	C_2H_4OH	C_2H_4OH	8.15	I.20	H	H	NO_2	H	C_2H_4OH	C_2H_4OH	3.14
I.7	H	Cl	H	H	C_2H_4OH	C_2H_4OH	7.83	II.1	-	-	-	H	C_2H_4CN	C_2H_4CN	4.95
I.8	H	H	H	H	C_2H_5	C_2H_4OH	7.27	II.2	H	-	-	H	C_2H_4OH	C_2H_4OH	12.71
I.9	H	H	H	CH_3	C_2H_4OH	C_2H_4OH	6.48	II.3	H	-	-	H	C_2H_4CN	C_2H_4CN	16.58
I.10	CN	H	H	H	C_2H_4OH	C_2H_4OH	6.46	II.4	OCH_3	-	-	H	C_2H_4OH	C_2H_4OH	14.23
I.11	H	NO_2	H	H	C_2H_4OH	C_2H_4OH	6.37	II.5	CH_3	-	-	H	C_2H_4OH	C_2H_4OH	15.26
I.12	OCH_3	H	H	H	C_2H_4OH	C_2H_4OH	6.14	II.6	-	-	-	H	C_2H_4OH	C_2H_4OH	18.87
I.13	H	H	H	H	H	C_2H_4OH	6.03	II.7	-	-	-	H	C_2H_4OH	C_2H_4OH	21.01
I.14	H	CH_3	H	H	C_2H_4OH	C_2H_4OH	5.94								

PLS calculations were performed by the SIMCA package [SIMCA-P+, version 12.0; Umetrics AB: Umeå, Sweden, http:www.umetrics.com]. The goodness of prediction was tested by the leave-7-out crossvalidation approach. In addition, the predictive power of the model was tested by the following statistical measures, too [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:

$$q^2 > 0.5 \quad (3)$$

$$R^2 > 0.6 \quad (4)$$

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

or

$$\frac{(R^2 - R_0'^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \le k' \le 1.15 \quad (6)$$

$$\left|R_0^2 - R_0'^2\right| < 0.3 \quad (7)$$

RESULTS AND DISCUSSIONS

In a previously published paper [5], MLR, MTD and CoMFA approaches were applied to a series to 27 dyes. A poorer correlation with the ClogP (the calculated octanol-water partition coefficient) parameter (r^2 = 0.32) and a good correlation with the MTD parameter (r^2 = 0.924) were obtained suggesting that steric interactions are more important in comparison to the hydrophobic ones. Comparative Molecular Field Analysis (CoMFA), gave r^2 = 0.925, and q^2 (cross-validated r^2) = 0.776 for 2 PCs (principal components), emphasizing same steric contribution for enhancing the dye affinity. In addition, correlation with a one-dimensional descriptor (the dye molecular length), derived from the 3D dye structures gave similar results to the CoMFA ones. It was concluded that steric fields are well approximated by molecular length, while electrostatic interactions appeared to be less important. The affinity of binding was found to be less specific in terms of pharmacophoric constraints.

In this paper the same series of 27 dyes was studied by molecular mechanics calculations and the optimized structures thus derived were used to calculate dye descriptors. PLS calculations were performed to correlate the dye affinity values with the calculated descriptors. A training set of 20 compounds and a test set of 5 compounds: I.10, I.12, I.13, I.18 and II.4 (table 1) were considered.

The test set compounds were selected consulting the scores (t_{f+1} function of t_f) scatter plots of the first five principal components for the principal component analysis model constructed using the X matrix of the set of descriptor variables included in the final PLS model for the 20 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.

Starting from the descriptor matrix containing all variables, following descriptors were found to be significant and were included in the final PLS model: nBM (number of multiple bonds), Ui (unsaturation index), nCIC (number of rings), SCBO (sum of conventional bond orders (H-depleted)), nAB (number of aromatic bonds), nS (number of Sulfur atoms), nR05 (number of 5-membered rings), nR09 (number of 9-membered rings), nThiazoles(number of Thiazoles), nCIR(number of circuits), nBO (number of non-H bonds), nCar (number of aromatic C(sp^2)), Hyper80 (Hypertens-80 (Ghose-Viswanadhan-Wendoloski antihypertensive-like index at 80%)), Neop80 (Neoplastic-80 (Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 80%)), nSK (number of non-H atoms), nCb- (number of substituted benzene C(sp2)), MW (molecular weight), ALOGP2 (Squared Ghose-Crippen octanol-water partition coeff. ($logP^2$)), AMR (molar refractivity), Mp (mean atomic polarizability (scaled on Carbon atom)), Mv (mean atomic van der Waals volume (scaled on Carbon atom)), ARR (aromatic ratio), MLOGP2 (Squared Moriguchi octanol-water partition coeff. ($logP^2$)), BLTF96 (Verhaar model of Fish base-line toxicity for Fish (96h) from MLOGP (mmol/l)), MLOGP (Moriguchi octanol-water partition coefficient), BLTA96 (Verhaar model of Algae base-line toxicity for Algae (96h) from MLOGP (mmol/l)), TPSA(Tot) (topological polar surface area using N, O, S, P polar contributions), BLTD48 (Verhaar model of Daphnia base-line toxicity for Daphnia (48h) from MLOGP (mmol/l)), nN (number of Nitrogen atoms), Depr50 (Depressant-50 (Ghose-Viswanadhan-Wendoloski antidepressant-like index at 50%)), Ss (sum of Kier-Hall electrotopological states), AMW (average molecular weight), Infec50 (Infective-50 (Ghose-Viswanadhan-Wendoloski antiinfective-like index at 50%)), Neop50 (Neoplastic-50 (Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 50%)).

In the final PLS model the regression coefficients were transformed as function of the original variables (figure 1). The importance of descriptors was evaluated by the VIP (Variable Influence on Projection) values [11], which summarizes the importance of the x variables in the model. VIP values higher than 1.0 were considered (figure 2). In figures 1 and 2 the error bars, which indicate 95% confidence intervals based on jackknifing, emphasize the certainty of the chosen variables.

Compounds no. I.20 (probably because of steric hindrance between the NO_2 group placed in ortho position with respect to the azo group) and II.1 (the disperse dye with smallest heterocyclic

moiety) (table 1) were found as outliers, based on the Hotelling T^2 criterion and were omitted from the final PLS model.

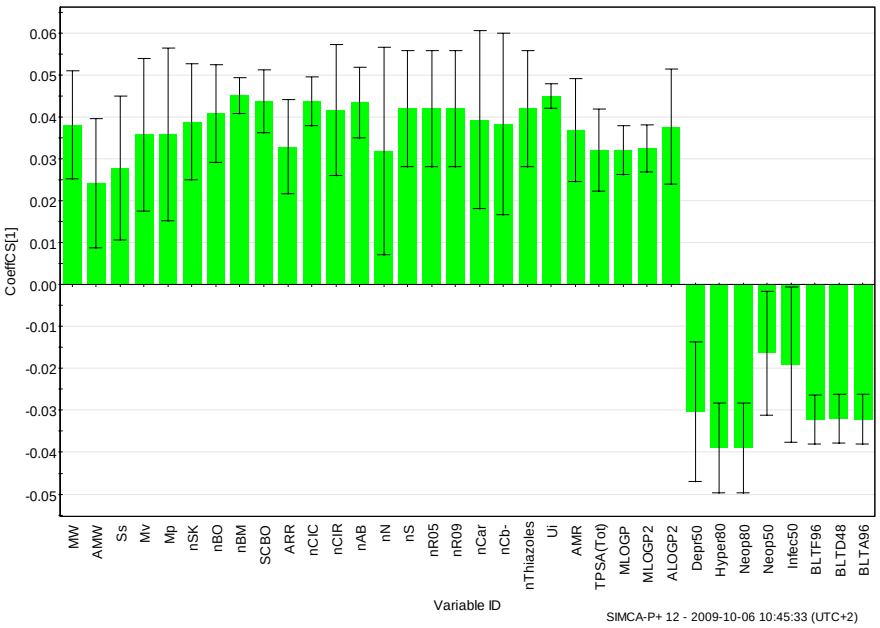


Figure 1. PLS regression coefficients transformed as function of the original variables after 1 principal component.

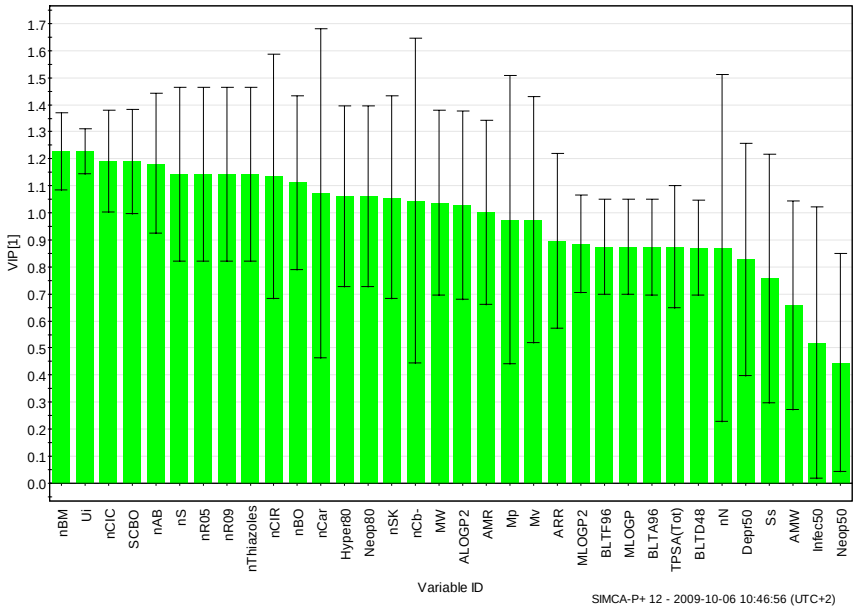


Figure 2. VIP values calculated for 1 principal component.

An acceptable PLS model with 1 principal component was obtained: $R^2X(cum) = 0.617$, $R^2Y(cum) = 0.959$, $Q^2(cum) = 0.953$, where $R^2Y(CUM)$ are the cumulative sum of squares of the entire Y's explained by all extracted principal components and $Q^2(CUM)$ is the fraction of the total

variation of the Y's that can be predicted for all the extracted principal components. The dependence between experimental and predicted affinity values for the training and test set is presented in figure 3.

The predictive power of the PLS model was then checked by the criteria stated in equations (3) to (7) [10]. All these calculated criteria indicated a model with predictive power, respectively:

$Q^2(CUM) = 0.953 > 0.5$

$R^2 = 0.934 > 0.6$

$\frac{(R^2 - R_0^2)}{R^2} = 0.014 < 0.1 \quad \text{and} \quad 0.85 \le k = 1.072 \le 1.15$

$\frac{(R^2 - R_0'^2)}{R^2} = 0.042 < 0.1 \quad \text{and} \quad 0.85 \le k' = 0.92 \le 1.15 \text{ and } \left|R_0^2 - R_0'^2\right| = 0.026 < 0.3 .$

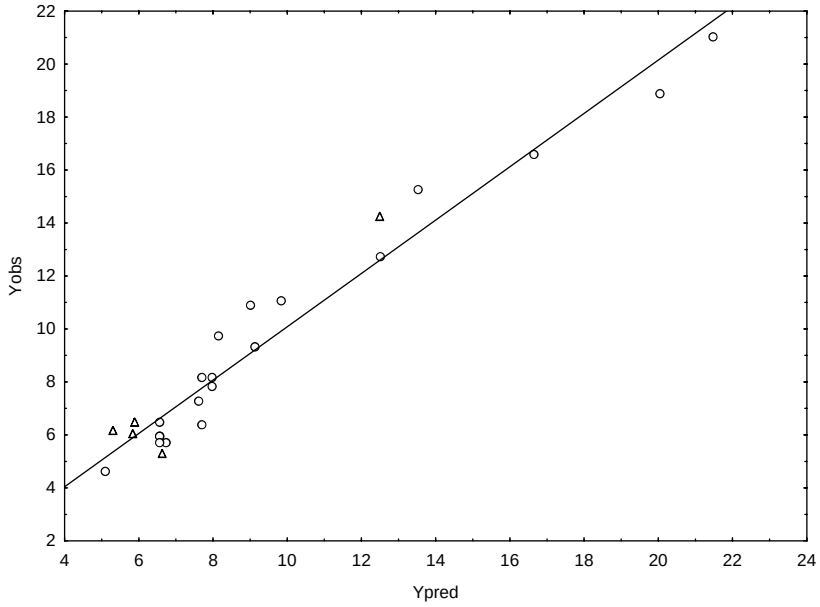


Figure 3. Experimental (Yobs) *versus* predicted (Ypred) affinity values of the final PLS model (training set marked by circles, test set marked by triangles).

Several dye features which influence their toxicity were derived from the VIP values. The presence in the dye molecules of heterocyclic thiazole moiety having many condensed phenyl rings and higher number of substituted benzene groups, as well as molecules with higher volumes and polarizabilities increased the dye affinity. Calculated drug-like ability of the dye molecules, expressed by indices, like: Ghose-Viswanadhan-Wendoloski antihypertensive-like index at 80% and Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 80% [12] unfavored their binding to cellulose. The presence in the final PLS model of the squared Ghose-Crippen octanol/water partition coefficient indicates the existence of optimum dye hydrophobicity for the cellulose affinity.

Hydrogen bonding is not present and the steric interactions are dominant in disperse dye-cellulose binding, in accordance to previous findings [2, 5].

In comparison to the statistical results obtained by MLR, MTD and CoMFA [5], the PLS results indicate a better goodness of fit and prediction for the disperse dye binding to cellulose. The three-dimensional CoMFA model gave more useful information on the specificity of dye-fibre interactions, in comparison to the PLS model obtained from 0D, 1D and 2D variables.

CONCLUSIONS

Dye binding to cellulose was studied by correlations of dye affinity values with structural descriptors by the partial least squares (PLS) method. Dye structures were modeled by molecular mechanics and structural variables were derived from the optimized structures.

Simple 0D, 1D and 2D descriptors enabled us to obtain a PLS model with good statistical results. The presence of heterocyclic fragment having many phenyl rings attached to the thiazole moiety, higher number of substituted benzene moieties are favorable for increased affinity. Hydrogen bonding is not characteristic for disperse adsorption on cellulose. The goodness of fit and prediction of this model was better in comparison to previously published MLR, MTD and CoMFA ones for the same series of dyes. The PLS model calculated from 0D, 1D and 2D variables did not give information in terms of specificity of dye-fibre interactions as in case of CoMFA.

REFERENCES

1. Peters, R. H., Textile Chemistry. The physical Chemistry of Dyeing. Vol. III, Elsevier Scientific Publ. Co., Amsterdam, 1975.

2. Shibusawa, T.; Uchida, T. Sen'i Sakkaishi 1986; 42: T84 - T91.

3. Timofei, S.; Kurunczi, L.; Schmidt, W.; Simon, Z. Dyes Pigm 1995; 29: 251-258.

4. Timofei, S.; Kurunczi, L.; Schmidt, W.; Simon, Z. Dyes Pigm 1996; 32: 25-42.

5. Oprea, T. I.; Kurunczi, L.; Timofei, S. Dyes Pigm 1997; 33: 41-64.

6. Seu, G.; Mura, L. Am Dyest Rep 1984; 43 – 44.

7. Wold, H. Partial Least Squares, in: 'Encyclopedia of Statistical Sciences'. (Kotz S.; Johnson N. L., Eds.), Vol. 6, Wiley, New York, 1985, p. 581-591.

8. Höskuldsson, A. J Chemometrics 1988; 2: 211-228.

9. Hellberg, S.; Wold, S.; Dunn III W.J.; Gasteiger, J.; Hutchings, M.G. Quant Struct.-Act Relat 1985; 4: 1-11.

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. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold S. Multi- and Megavariate Data Analysis. Principles and Applications, Umetrics AB, Umeå, Sweden, 2001, p. 94-97
12. Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. J Comb Chem. 1999; 1: 55-68.