Dyes and Pigments 153 (2018) 35–43

Contents lists available at ScienceDirect

Dyes and Pigments

journal homepage: www.elsevier.com/locate/dyepig

Hologram quantitative structure–activity relationship and topomer comparative molecular-field analysis to predict the affinities of azo dyes for cellulose fibers

Shuling Yu[a], Qiaoqiao Zhou[a], Xuliang Zhang[a], Shuxin Jia[a], Ying Gan[a], Yahong Zhang[a], Jiahua Shi[a], Jintao Yuan[b,*]

[a] Key Laboratory of Natural Medicine and Immune-Engineering of Henan Province, Henan University, Kaifeng, Henan, 475004, PR China
[b] College of Public Health, Zhengzhou University, Zhengzhou, 450001, PR China

ARTICLE INFO

Keywords:
Hologram quantitative structure–activity relationship
Topomer comparative molecular-field analysis
Azo dyes

ABSTRACT

A hologram quantitative structure–activity relationship (HQSAR) and topomer comparative molecular-field analysis (CoMFA) were performed on 51 molecules that belong to the azo dyes to determine their affinities to cellulose fiber. The best HQSAR model was obtained by using atoms. Hydrogen atoms as a fragment distinction and a fragment size of 3–6 showed a leave-one-out cross-validated correlation coefficient (q^2_{LOO}) of 0.849 and a non-cross-validated correlation coefficient (r^2) of 0.927. The best topomer CoMFA model with steric and electrostatic field parameters based on three fragments gave satisfactory results (q^2_{LOO} = 0.821; r^2 = 0.934). External predictions were made on a test set and compared with previously reported models. Topomer CoMFA steric, electrostatic, and HQSAR atomic contribution maps were generated to analyze the structural features of this data set that govern their affinity potency.

1. Introduction

The dye industry produces thousands of dyes, which represent an abundant class of colored organic compounds. Affinity is one of the major properties of dyes. Theoretical studies on the mechanism of dye adsorption have been carried out by different modern methods, but there is no single interpretation of dyeing theory [1]. The field of quantitative structure–activity relationships (QSAR) constitutes an important tool to study dye–fiber interactions [2]. The advantages of this approach to dye adsorption on cellulose fiber are related either to a description of the mechanisms at a molecular level or to the predictability of the proposed models, which can lead to the design of new dyes with higher affinities for cellulose fiber.

Azo compounds have vivid colors, comprise approximately two-thirds of all synthetic dyes, and are the most widely used and structurally diverse class of commercial organic dyes [3]. Therefore, the study of dye-fiber affinity is important. In recent years, some QSAR models have been developed to correlate azo dye affinity with the dye molecular structure. Funar–Timofei and Schüürmann [4] reported QSAR models for a series of 30 anionic azo dyes by comparative molecular-field analysis (CoMFA) combined with a semiempirical quantum chemical AM1 calculation. Later, they investigated the 30 dyes again [5] by using CoMFA and solution-phase molecular descriptors. The same 30-anionic-azo-dye dataset was studied by Polanski et al. [6] by using comparative molecular surface analysis, by Zhokhova et al. [7] by using multiple linear regression (MLR) based on fragment descriptors that were calculated by using the NASAWIN software package, and by Bak and Polanski [8] by using the receptor-independent four dimensional (4D)-QSAR method that is based on a self-organizing mapping approach. The affinities of 21 heterocyclic monoazo dyes were modeled by Funar–Timofe et al. [9] by using MLR, CoMFA, and a comparative molecular similarity index analysis, and by Bak et al. [10] by using receptor-dependent 4D-QSAR methods. More recently, Wang et al. [11] combined 30 anionic azo dye datasets and 21 heterocyclic monoazo dye datasets to develop MLR and artificial neural network (ANN) models based on descriptors that were calculated by using Dragon software. Although the MLR and ANN models that were reported by Wang et al. [11] provided a good fit, and were based on more azo dye molecules than previous reports, the information that was encoded in the descriptors that were included in the MLR and ANN models could not provide intuitive graphical results to guide the design of the new azo dyes. Therefore, the establishment of simple, repeatable, and intuitive models for azo dyes is important.

Holographic QSAR (HQSAR) is an extraordinary QSAR technique

* Corresponding author.
E-mail address: jtyuan@zzu.edu.cn (J. Yuan).

https://doi.org/10.1016/j.dyepig.2018.01.053
Received 13 December 2017; Received in revised form 29 January 2018; Accepted 30 January 2018
Available online 06 February 2018
0143-7208/

[12–14]. In this method, an extended form molecular hologram is used as a molecular descriptor to fingerprint encode all possible molecular fragments, including linear, branched, cyclic, and overlapping molecular features. HQSAR avoids the need for a three-dimensional (3D) structure, putative binding conformations, and molecular alignment in CoMFA, and also averts the selection and calculation or measurement of the physicochemical descriptors that are required by classical QSAR [15,16]. Topomer CoMFA is a second-generation CoMFA method. An obvious advantage of topomer CoMFA compared with traditional CoMFA is that topomer CoMFA models can be generated by splitting the molecules into R-groups, aligning each fragment, and calculating the steric and electrostatic field descriptor values automatically. A topomer CoMFA prediction is almost entirely objective, and depends only on the two-dimensional (2D) connectivity: user-specified fragmentations, and measured activities of the training set structures [17–19]. Both methods can provide intuitive graphical results. Therefore, HQSAR and topomer CoMFA were used in this study because of the aforementioned advantages.

This work aims to identify the stereo-electronic requirements of azo dyes for an increased affinity to cellulose fiber. HQSAR and topomer CoMFA analysis were performed on the previously reported 51 azo dyes to reveal the structural factors that govern their affinity. The statistical parameters of the HQSAR and topomer CoMFA models were compared with previously reported models. Finally, the obtained results from the HQSAR and topomer CoMFA were used as a guide to design new azo dyes.

2. Materials and methods

2.1. Data set

Fifty-one azo dyes were collected from literature [11]. The compounds and their corresponding experimental dye affinity values are listed in Table 1. The entire data set was divided into training and test sets by the DUPLEX algorithm [20], and the distribution of training and test sets was identical to that in literature [11] to facilitate a comparison of results. The data set of 51 dyes was divided into training (41 dyes) and test (10 dyes) sets. The structures of the 51 dyes, and the affinity values and distributions of the training and test sets are provided in Table 1. For the topomer CoMFA study, the 3D structures of the 51 azo dyes were constructed using Sybyl-x 2.0 software. Each molecule energy in the data set was minimized through the gradient-descent method and by using a Tripos force field and Gasteiger–Huckel charges.

2.2. HQSAR

In general, HQSAR analysis includes three main steps: (1) generating sub-structural fragments for each molecule in the data set, (2) encoding these fragments in holograms, and (3) correlating the latter with the available property value. In HQSAR, the input molecule is broken into a series of unique structural fragments (linear, branched, and overlapping). The information that is contained in each fragment is defined by six fragment distinction parameters, as follows: the atomic number (A), the bond type (B), the atomic connection (C), hydrogen (H), chirality (Ch), and donor/acceptor (D). Each fragment is assigned a specific integer from 0 to 231 via a cyclic redundancy check algorithm. Every integer corresponds to a bin in an integer array of fixed-length hologram length (which is usually one of the 12 prime numbers that range from 53 to 401), in which all elements are set to zero initially. Bin occupancies are the descriptors of HQSAR modeling [12], and the partial least square (PLS) method was used to develop the QSAR model. HQSAR models can be displayed graphically as color-coded structure diagrams, in which the color of each atom reflects its contribution to potency variation. The red and green ends of the spectrum reflect the negative and positive contributions, respectively. Atoms with intermediate contributions are colored white [21]. HQSAR modeling was performed with Sybyl-x 2.0 software.

2.3. Topomer CoMFA

The topomer CoMFA methodology combines CoMFA with topomer technology to overcome the alignment problem of CoMFA. A topomer is a molecular fragment, which is prescribed by its conformation and its position [22], and indicates any structure that contains at least one open valence. Similar to CoMFA, the steric field and electrostatic field were calculated using the standard Tripos force field method and sp3 carbon probe atom with a +1 charge. Topomer CoMFA models are built using PLS, and steric and electrostatic descriptors were used as independent variables. The optimum number of latent variables (LVs) is evaluated by leave-one-out (LOO) cross–validation. Finally, the field contour maps are generated by using the StDev*Coeff field type.

2.4. Model evaluation

The goodness of fit and prediction were used to evaluate the model performance. The goodness of fit was examined by the coefficient of determination (r^2), non-cross validated standard error (SEE), and Fisher's test (F) value. Cross-validation (CV) is the most commonly used technique for internal validation. The outcome from the CV procedure is cross-validated r^2 (q^2), which is used as a criterion of robustness and predictive ability of the model. The optimum numbers of LVs of models were determined by the highest LOO q^2 (q^2_{LOO}). Compared with the LOO procedure, the leave-many-out (LMO) CV allows more chemicals to be omitted as a prediction to test the model's stability. Thus, the cross-validated r^2 (q^2_{LMO}) of LMO CV accounts for more model extrapolation than q^2_{LOO} [12,23]. The selection of good predictive models depends on q^2_{LMO}. In this study, LMO CV was performed by using five groups, and the procedure was repeated 10 times. The mean of 10 readings was $q^2_{\mathrm{LMO\text{-}5}}$. The goodness of the model prediction depends on the challenging test set prediction that is reflected by its predictive r^2. Several external validation parameters (Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and the concordance correlation coefficient (*CCC*)) were proposed in literature [24] for the external validation of QSAR models were employed in the HQSAR and topomer CoMFA analyses.

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2} \quad (1)$$

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2} \quad (2)$$

$$Q^2_{F3} = 1 - \frac{[\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2]/n_{EXT}}{[\sum_{i=1}^{n_{TR}} (y_i - \bar{y}_{TR})^2]/n_{TR}} \quad (3)$$

$$CCC = \frac{2\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}} (y_i - \bar{y})^2 + \sum_{i=1}^{n_{EXT}} (\hat{y}_i - \bar{\hat{y}})^2 + n_{EXT}(\bar{y} - \bar{\hat{y}})^2} \quad (4)$$

where, y_i is the experimental data values, $\hat{y}_i$ is the predicted value, $\bar{y}_{TR}$ is the mean value in the training set, the symbol TR and EXT are used for training and test sets, respectively, $\bar{y}$ is the average of the experimental data values, $\bar{\hat{y}}$ is the average of the predicted data values, n is the number of compounds.

3. Results and discussion

3.1. HQSAR analysis

Several parameters related to hologram generation, such as fragment size, and fragment distinction, affect the HQSAR model quality. To derive the best HQSAR model, the influence of various combinations of parameters on the HQSAR model was studied. First, HQSAR analysis was conducted using default parameters, including the six default prime