Available online at www.sciencedirect.com
SCIENCE DIRECT®
Dyes and Pigments 62 (2004) 61–76
DYES and PIGMENTS
www.elsevier.com/locate/dyepig

# Mapping dye pharmacophores by the Comparative Molecular Surface Analysis (CoMSA): application to heterocyclic monoazo dyes

Jaroslaw Polanski[a,*], Rafal Gieleciak[a], Miroslaw Wyszomirski[b]

[a]*Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland*
[b]*University of Bielsko-Biala, PL-43-310 Bielsko-Biala, Poland*

Received 1 August 2003; received in revised form 14 November 2003; accepted 16 November 2003

**Abstract**

Despite recent investigations aimed at developing for dye molecules Quantitative Structure Activity Relationships on the basis of the data that describe 3D molecular structure of these molecules (so-called 3D QSARs), it is still uncertain whether a pharamacophore hypothesis can be used for modeling dye–cellulose affinities. We have used the Comparative Molecular Surface Analysis (CoMSA) method for modeling the cellulose affinity of heterocyclic monoazo dyes. We used the molecule of the highest cellulose affinity as the template pattern. Moreover, we also used so-called alternative mapping strategy with smaller molecular fragments applied as different templates to investigate the contribution of the individual molecular moieties to the compound–cellulose attraction and for an estimation of the relative importance of the steric and electrostatic interactions. We obtained models significantly outperforming those reported in Comparative Molecular Field Analysis studies. Our results proved the previous hypothesis that polar interactions are decisive for the dye–fiber affinity. Moreover, the highly predictive models that were obtained indicate that a pharamacophore (or more precisely *tinctophore)* concept can be used efficiently for the description of the dye–fiber interactions.


*Keywords:* Comparative Molecular Surface Analysis; Heterocyclic monoazo dyes; Kohonen neural networks; Tinctophore

## 1. Introduction

Many different mechanisms can be involved in the interactions of chemical molecules with their environment that determine the molecular effects evoked. Modeling environmental response is one of the basic techniques available to understand and explain these mechanisms. The comparison of a series of molecules that stimulate the environment can provide a useful information on the environment allowing us also for the optimization of the stimulant structure. In particular in drug design, modeling environmental response on the basis of such a comparison is a popular method for the investigations of the drug–receptor interactions. A term Quantitative Structure Activity Relationships (QSAR) refers to such investigations. Recent developments in computational chemistry allows us also to compare 3D molecular structures and consequently to investigate 3D QSARs. In pharmacy a pharmacophore concept stands

* Corresponding author.
*E-mail address:* polanski@us.edu.pl (J. Polanski).

0143-7208/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dyepig.2003.11.003

behind this strategy. The notion of pharmacophore is based on the attempting to identify a certain subset of the atoms that forms a common or similar pattern for all active molecules. In fact, we do not have any information on what the relation between a real receptor and this subset is. Pharmacophore mapping is a broad technique that includes a variety of experimental and computational approaches. This technique brings together different methods including synthesis of the diversified molecular structures in the attempt to measure chemical, physical and biological characteristics and to describe structure–activity relationships (SAR) and a variety of techniques used for further computational analysis of these data. A focus on the molecule that stimulates the environment or binds the receptor, i.e., the effector or ligand molecule, is a common feature of this strategy. Actually, it is not only this molecule that decides molecular effects. In fact a molecule interacting with the environment forms a complex system which very often cannot be described doing without the corresponding information on the environment. Consequently, the molecular environment that complements an effector molecule is the important counterpart limiting the efficiency of such modeling.

The interaction between a dye molecule and cellulose is a complicated phenomenon, which can be described by the Langmuir isotherm [1,2]. Most adsorption isotherms correspond to Freundlich or Langmuir equations and the applicability of the Langmnuir isotherms for describing cellulose dyeing suggests specific interactions between dye and fiber [2]. An isotherm does not, however, provide a molecular description of the process. Moreover, we cannot use such an approach for the optimization of the molecular structure of dye. The influence of the electrostatic, van der Waals or hydrogen bonding as well as hydrophobic forces on dyeing process has been investigated. Compare Ref. 2 for a brief review. On the other hand, it has been speculated that specific binding sites exist on the crystalline region of the supramolecular cellulose structure that forms holes and cavities capable of incorporating a dye molecule [3]. Does this mean that a similarity between the drug–receptor and dye–fiber interactions makes possible to extend a pharmacophore concept and develop an idea of *tinctophore* in dye chemistry to predict tinctorial properties of dyes by the use of QSAR or related methods? Although it is not clear if we can treat it similarly to the contacts taking place during targeting a receptor by a drug molecule, several QSAR studies have been published recently [4–11] that make use of this concept in investigations of cellulose dyeing. Both 2D and 3D QSAR modeling have been applied including the Hansch, MTD and Comparative Molecular Surface Analysis (C0MFA) methods that appeared to provide quite satisfactory models for different compound series [8–10,12]. In particular, the results of the CoMFA method indicated that the electrostatic field dominantly contributes to the dyeing affinity. On the other hand, dye–cellulose interactions seemed to be less *specific* than drug–receptor interactions [2]. The specificity of molecular recognition and binding indicates that receptors formed by proteins in a form of cavities tightly fit ligands or drugs. It means that only strictly defined or specific molecules are strongly favored by a given receptor. In our previous publication we investigated the dye–fiber affinities for the anionic azo and anthraquinone dyes [13]. In spite of some substantial differences between dye *tinctophores* and drug pharmacophores 3D QSAR strategies can provide an efficient tool in such cases. The aim of the current study is a systematic analysis of the tinctorial properties of heterocyclic monoazo dyes using the Comparative Molecular Surface Analysis (CoMSA) as a novel method for 3D QSAR modeling [14–17]. Dyeing cellulose with these compounds has been investigated by Alberti et al. [18–20] and CoMFA study has been reported elsewhere [12]. We endeavored to use the CoMSA method both for verifying the applicability of the pharmcophore concept in dye chemistry as well as for discussing some more general rules in modeling 3D QSAR.

## 2. Methods

### 2.1. Self-organizing neural networks

A self-organizing neural network [21] is an unsupervised architecture. This network includes only a single layer, usually a 2D grid of neurons.

The 2D topology of the grid implies that we can distinguish the neighborhood relations between the nodes by defining the distances between them. N-dimensional data vectors presented into such a network are distributed between these neurons in such a way that those that are similar are put closer (into the neurons that are closer neighbors) to each other than those that differ. During so-called learning each $N$-dimensional input vector is compared with the $n$-element ($n=N$) weight vectors describing each neuron to detect the one into which the individual input vector will be projected, At the beginning of the learning the weight vectors are set randomly (in newer versions more effective strategies have been developed) and the network is presented with the first input vector. A formal criterion for the selection of the winner ($\mathrm{out_c}$) can be based for example on the Euclidean distance between a vector ($x$) and a weight ($w$).

$$\mathrm{out_c} \leftarrow \min\left[\sum_{i-1}^{m}\left(x_{si}-w_{ji}\right)^2\right] \quad (1)$$

Then the weight of the winner is corrected to decrease this distance. This means the neuron better recognizes this individual input. Moreover, the weights of the neighboring neurons are also corrected to attract similar inputs. Now the network is presented with the second vector and so on.

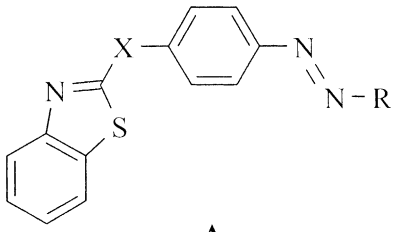
Self-organizing neural network (SOM) is a technique designed to reduce the dimensionality of the data while preserving topology. This is also an efficient clustering and data compression tool. Eventually, it can also be used for visualization purposes. Bioinformatics is one of the recent implementations that illustrate the importance of this technique [22–24]. The method has also been applied in chemistry [25,26], in particular, for 2D mapping of the electrostatic potential of 3D molecular surfaces [26,27] or partial atomic charges within a molecule [28]. The ability to compress the data and to later reconstruct the 3D object from the 2D representation makes this procedure an interesting tool for molecular design [25,26,29]. Such maps have been used for the visualization of the interactions of individual molecules with biological receptors or drug design [25,26]. It can also be a useful tool for pharmacophore mapping [30].

### 2.2. Comparative Kohonen mapping and CoMSA

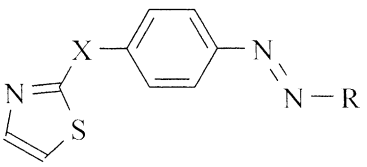
The different size of the molecular objects is one of the main obstacles making problems while comparing their 3D structure. If we for example would like to use the method described above for a comparison of the surfaces of two molecules of two different sizes, then, a single Kohonen neuron within each map will correspond to the different area on the surface of each molecule. Any quantitative comparison of such maps would be very inaccurate if the molecules significantly differ. On the other hand, if we prepare the maps of the different sizes to include the area of the similar size in a single neuron we cannot easily compare these maps. The reason for that is a fact that in computational approaches the objects are described by vectors or matrices that are later manipulated to produce a value characterizing a degree of similarity. Usually, such a manipulation needs, however, the same size of a matrix or vector.

In this particular case the Kohonen network can be used as an instrument that helps us to achieve this. Thus, the Kohonen network that learned the input data can be used to process the data characterizing any additional input object. Because we are using only a single network the output map is always of the same size, the network itself normalizes the dimensionality of the data. The precise definition and 3D representation of molecular surface are not easy to derive. Moreover, the presentation of molecular surfaces as 2D images always needs some projection or data transformation in order to reduce the dimension (from 3D to 2D) of the data to be visualized. Gasteiger and Zupan pioneered in the use of the Kohonen neural scheme for this purpose [26]. They illustrated the molecular electrostatic potential in a form of 2D map. The Cartesian coordinates ($x$,$y$,$z$) of the points sampled at the molecular surface form an input to Kohonen network, that is later used for the projection of the molecular electrostatic potential. Using the knowledge of the template molecule learnt by the network we can design a precise scheme for the comparison of molecules. This scheme uses a single Kohonen network (a template network) for processing data from more than one molecule. Then, we can sample points on

Table 1
Experimental dye affinity ($-\Delta\mu^0$) [12] of heterocyclic azo dyes arid this included by the CoMSA method



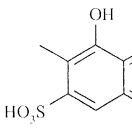

A



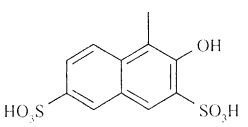

B

| No. | | X | R | $-\Delta\mu^0$ (kJ/mol)[a] | $-\Delta\mu^0$ [b] | $-\Delta\mu^0$ [c] |
|---|---|---|---|---|---|---|
| 1. | A | –NIH– | A | 6.78 | 7.73 | [d] |
| 2. | A | –NH– | B | 9.20 | 8.94 | 8.45 |
| 3. | A | –NH– | D | 12.60 | 11.96 | [d] |
| 4. | A | –NIH– | E | 15.30 | 15.29 | 14.66 |
| 5. | A | O | A | 3.26 | 2.47 | [d] |
| 6. | A | O | B | 5.27 | 5.29 | 4.60 |
| 7. | A | O | D | 7.61 | 8.30 | [d] |
| 8. | A | O | E | 10.30 | 10.49 | 10.38 |
| 9. | A | G | 0 | 10.20 | 10.29 | [d] |
| 10. | A | S | A | 1.26 | 0.57 | 1.48 |
| 11. | A | S | B | *3.56* | 3.81 | [d] |
| 12. | A | S | D | 5.02 | 5.76 | 6.04 |
| 13. | A | S | E | 8.45 | 8.69 | [d] |
| 14. | A | S | G | 8.12 | 8.44 | 8.25 |
| 15. | B | –NH– | E | 15.33 | 14.82 | [d] |
| 16. | B | –NH– | D | 12.60 | 11.96 | 12.99 |
| 17. | B | –NH– | B | 9.24 | 9.01 | [d] |
| 18. | B | –NH– | A | 6.80 | 7.50 | 7.62 |
| 19. | A | S | C | 5.86 | 7.21 | [d] |
| 20. | A | S | F | 10.33 | 8.77 | 9.14 |
| 21. | A | S | E (acid coupled) | 9.75 | 9.25 | [d] |

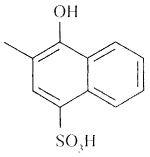
[a] Experimental affinities acc. to Ref. [12].
[b] CoMSA (size of Kohonen maps: 25×25; MD = 1.0; *en*).
[c] External predictions, details in text.
[d] Training series, details in text.



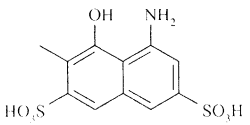

A



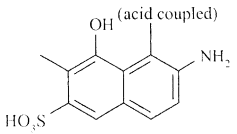

B



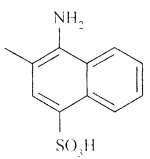

C



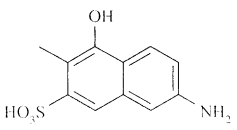

D




E




F




G

the molecular surface of a template molecule. A three-element ($x$,$y$,$z$) vector represents each point. Thus, a Kohonen network with three inputs can learn a topology of the template data. Further such a network can process similar data describing any particular molecule, irrespective of the molecule size yielding a map of the same size as that for the template molecule. Term "comparative mapping" or "template approach" was used to determine such a procedure. In our previous publications we described the use of the Kohonen neural network in QSAR investigations. In particular, we designed a 3D QSAR method by a coupled neural network and PLS system called Comparative Molecular Surface Analysis (CoMSA) [14–17,30,31]. Improvements to the CoMSA have also been published by Hasegawa et al. [32].

## 3. Experimental

### 3.1. Model building

All the experimental data given in Table 1, i.e., the dye affinities $\Delta\mu^{o}$ of heterocyclic monoazo dyes are from Ref. 12 that reports the CoMFA study of the compounds investigated previously by Alberti et al. for dyeing cellulose [18–20]. We used Gesteiger's software package for modeling purposes. The 3D coordinates of all molecules were obtained from the 3D structure generator CORINA [33–35] Partial atomic charges were calculated by the PEOE method [36,37] and the SURFACE program was used for the calculation of the Coulomb electrostatic potential on the molecular surface.

### 3.2. Data analysis

#### 3.2.1. Kohonen mapping

The competitive Kohonen strategy [21] was used to construct a 2D topographic map from the signals of points sampled randomly at the molecular surface. As molecular surfaces are continuous the plane of projection was also selected to be a continuous surface. Thus we used a torus for this purpose, which was cut along two perpendicular lines and then spread into a plane, as shown in Fig. 1. Each neuron, $j$, was then defined by three weights, $w_{ji}$. The competitive training of the network was based on the rule that each point, $s$, of the molecular surface was projected into that neuron, c, that has weights, $w_{ci}$, that come closest to the Cartesian coordinates, $x_{si}$, of this point, s [Eq. (1)].

$$\mathrm{out_c} \leftarrow \min\left[\sum_{i=1}^{m}\left(x_{si}-w_{ji}\right)^2\right]$$

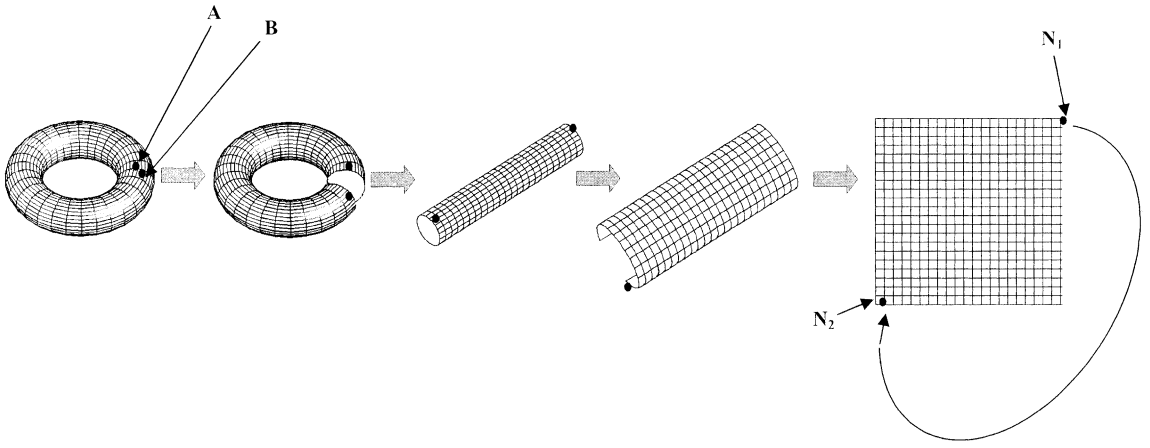



Fig. 1. The projection scheme based on torus (a) allows one to preserve topology during projection from the 3D to 2D map (b). The number of neighboring neurons in both representations is always the same. Two points A and B on torus (a) are projected to neurons N1 and N2 (b) that are considered as neighbors, respectively. Similarly, to 3D representation such a 2D map does not have any borders.

A projection of the electrostatic potential value (MEP) from the surface points, *s*, into such a 2D arrangement of neurons, after calculating the average MEP value within this particular neuron and scaling this values into the respective colors, results in the so called "feature map".

### 3.3. Comparative Kohonen mapping

A feature map illustrates the (MEP) of a single molecule. As, the weights of the Kohonen network contain an information on the shape of the particular molecular surface, the network can be used to compare molecular surface geometries of other molecules. The trained Kohonen network processes the signals from the surface of other molecule(s), i.e., the electrostatic potential of each input vector is projected through the network to obtain a series of comparative maps both for the template molecule and each analyzed molecule. The electrostatic potential values from the surfaces of the processed molecules are then projected into such a network allowing us to compare these parts of the molecule surfaces that can be superimposed. By a superimposition we mean an operation of the pairwise coverage of the atoms indicated in a pair of molecules. If we compare similar molecules or molecular fragments their molecular surfaces overlay occupying similar areas in the space. If we cover two different molecules their surfaces cannot be superimposed. Consequently, if we cannot superimpose a molecule (or its part) on the reference molecule (template) then the respective output neurons get no signal from this molecule. This results in the appearance of empty neurons that are coded by zeros. In Fig. 2, we illustrated this method by the comparative patterns of the monoazo dye series analyzed in the current publication.

We processed the electrostatic potential matrices (illustrated by color maps shown in Fig. 2b) to prepare different matrix types, as shown in Fig. 3. The original matrices that contain all elements including empty neurons form the first group, namely en matrices (*en*), as given in Fig. 3a. Next we transformed all negative values to $-1$ and all positive to $+1$ to obtain a new matrix type that we called *en*$(-1,0,+1)$ (Fig. 3b). Alternatively, all nonzero elements take a value of 1 in the *en*(0,1) matrices (Fig. 3c). Finally, we eliminated these columns within comparative patterns that include any empty neuron to obtain "non-empty-neurons" matrices (*nen*) (Fig. 3d). Thus elements that are represented by 0 in any of the original matrices in the series are eliminated in each matrix of the series. Conversely, the next type of the electrostatic potential matrices, non–nen matrices (*non–nen*), shown in Fig. 3e, are those that include only such elements which have at least one empty neuron in any matrix of the series. The inclusion of empty neurons, illustrating the geometrical differences of the molecules, should point to steric influences while their elimination overestimates electrostatic interactions. On the other hand, the substitution of the particular electrostatic potential values by a discrete $+1,-1$ should decrease the importance of electrostatic influences to the advantage of the steric influences.

All the molecules were superimposed before the calculation of the molecular surfaces. Covering each atom of the template molecule with the respective atoms of the analyzed molecule is performed to achieve molecular superimposition, as shown in Fig. 4. In practice, we used Match3D program [38] to carry out this operation. The KMAP 3.0 program [38] was used for the simulation of the Kohonen networks. The size of these networks was varied from $10\times10$ to $30\times30$. The output from this program was used for the calculation of the mean electrostatic potential values within each neuron and respective feature maps were transformed to a respective $N^2$ element vector, where $N$ is the number of neurons forming in the Kohonen map.

#### 3.3.1. PLS analysis

The PLS procedures were programmed within the MATLAB environment (MATLAB). As required by PLS method, the model was constructed for the centered data, i.e., the mean value for all variables were adjusted to take a value of 0. Its complexity was estimated based on the leave-one-out (LOO) cross-validation procedure (CV). The data was recentered (but not rescaled) for each crossvalidation run. In the LOO-CV one repeats the calibration $m$ times, each time treating the $i$th left-out object as the prediction object. The dependent variable for each left-out object is

calculated based on the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with $j$ factors is defined as:

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum(\mathrm{obs}_i - \mathrm{pred}_{i,j})^2}{m}} \quad (2)$$

where obs denotes the assayed value; pred—predicted value of dependent variable and $i$ refers to the object index, which ranges from 1 to $m$. Model with $k$ factors, for which RMSECV reaches a minimum, is considered as an optimal one.

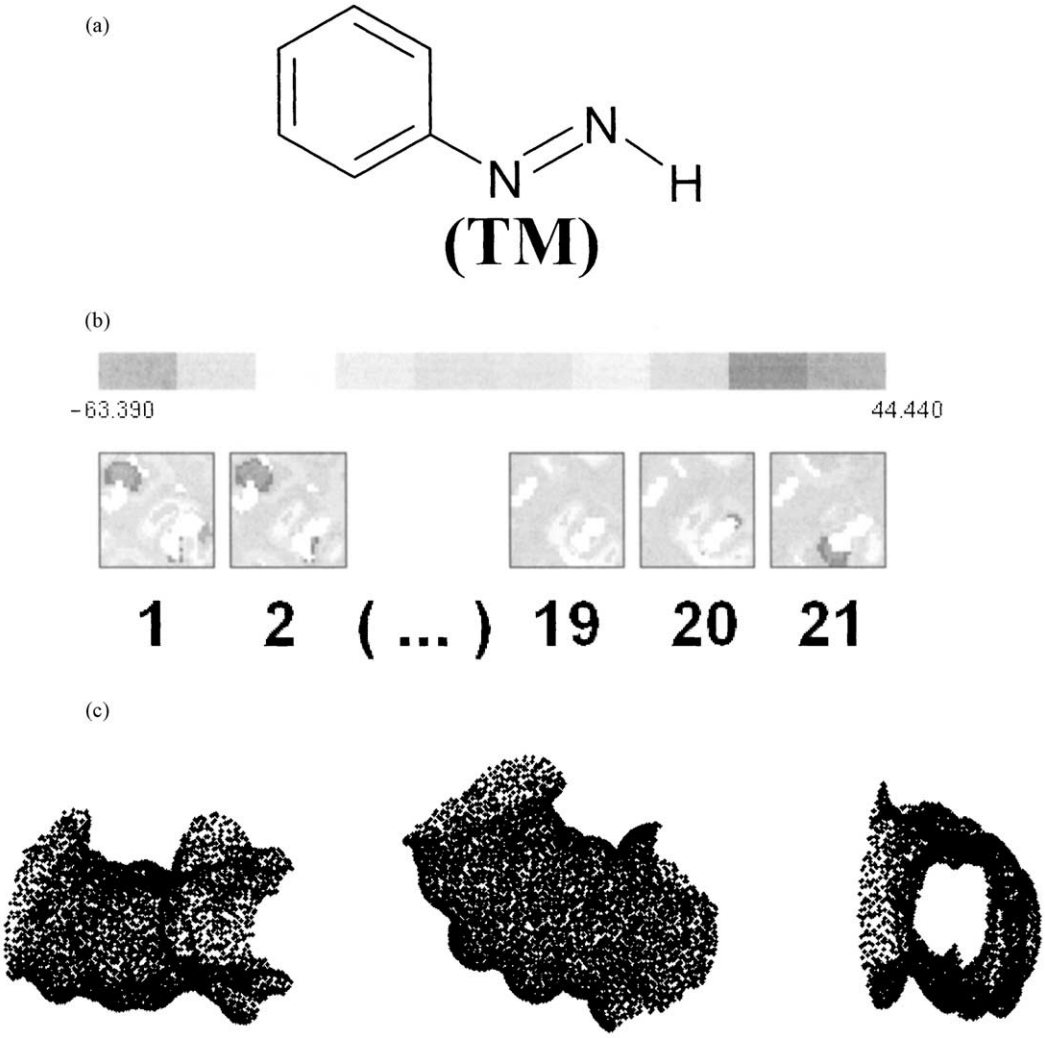



Fig. 2. Illustration of the Comparative Molecular Surface Analysis scheme. The template Kohonen network trained with the data from the arbitrarily template molecule TM (a) processes the data describing each molecule within the dye series **1–21** to obtain a series of the comparative electrostatic potential maps **1–21** (b). These patterns form a kind of superimposition plots. The backprojection of the data from such 2D maps to the surface of TM clearly shows this surface zones that form a common motif for the TM and individual molecules within the series. Three different orientations are shown (c). In particular, if we compare the surface of any dye molecule to the TM one, two areas with empty neurons appeared (a). The differential surface visualization (c) allows observing this effect in three dimensions. Two "wholes" in the surface resulted allows one to "look through" the TM molecule, as shown in the right most orientation.

We used the performance metrics that are accepted and widely used in CoMFA analyses, i.e., cross-validated $q^2$

$$q^2 = 1 - \frac{\sum(\text{obs}_i - \text{pred}_i)^2}{\sum(\text{obs}_i - \text{mean(obs)})^2} \quad (3)$$

where: obs—the assayed values; pred—predicted values, mean—mean value of obs and $i$ refers to the object index, which ranges from 1 to $m$; and cross-validated standard error $s$

$$s = \sqrt{\frac{\sum(\text{obs}_i - \text{pred}_i)^2}{m - k - 1}} \quad (4)$$

where: $m$—number of objects, $k$—number of PLS factors in the model.

The quality of external predictions was measured by the SDEP parameter:

$$\text{SDEP} = \sqrt{\frac{\sum(\text{pred}_i - \text{obs}_i)^2}{n}} \quad (5)$$

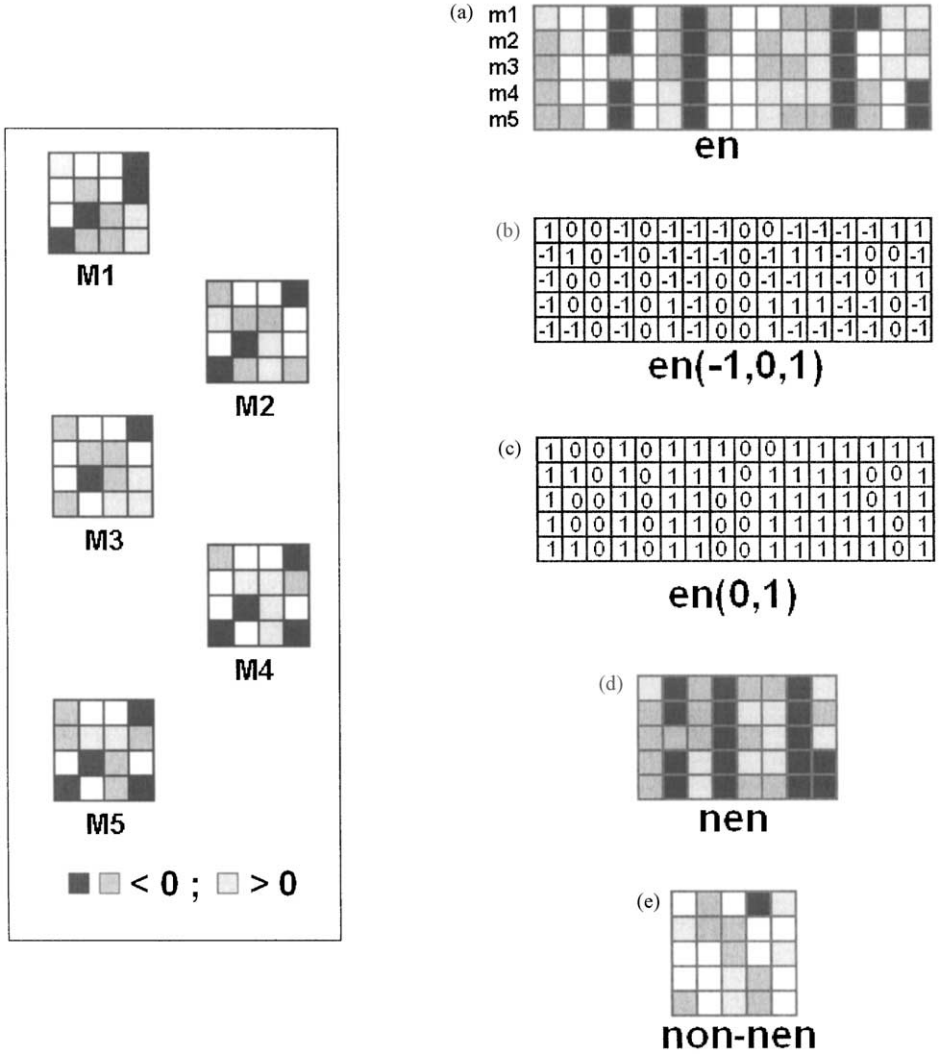



Fig. 3. The electrostatic potential matrices resulted from the CoMSA procedure and different protocols for their transformation. Details in text.

where: pred—predicted value, obs.—observed value, $n$—the number of compounds included in the test series.

## 4. Results and discussion

Table 1 and Fig. 5 illustrate the results obtained while using the CoMSA for modeling dye–fiber affinity. Thus, in the column 6 of Table 1 we specified the predicted cross-validated (LOO) values of the binding affinity according to the best model obtained. Fig. 5 compares different CoMSA models obtained. In this example molecule **15** of the highest dye–cellulose affinity was selected as the template for training the network. We checked three different superimposition modes SM1–SM3, which are indicated by encircling the atoms that were covered before the molecular data describing all compounds were processed sequentially by the comparative Kohonen network. The bar plots shown in Fig. 5 analyze the $q^2$ performances of CoMSA modeling. This indicates that model quality only slightly depends upon the superimposition mode. On the other hand, all the models are very predictive and $q^2$ ranges from 0.869 to 0.981. In Fig. 6 we analyzed a 3D patterns of the molecular superimpositions SM1–SM3 applied before conducting comparative Kohonen mapping. This shows that a perfect coverage of all compounds by each atom cannot be accomplished for the whole series. To further verify the influence of molecular superimposition we also applied an alternative mapping strategy [30,39], summarized in Fig. 7. Here the series of templates (T1–T3) used for training Kohonen network were selected among the motifs constructing the dye molecules. This allows us not only to observe the contribution of the different structural motifs for the celluose–dye affinity, but also simulates a situation in which molecules will be flexible enough to be covered by all the motifs indicated as the T1, T2

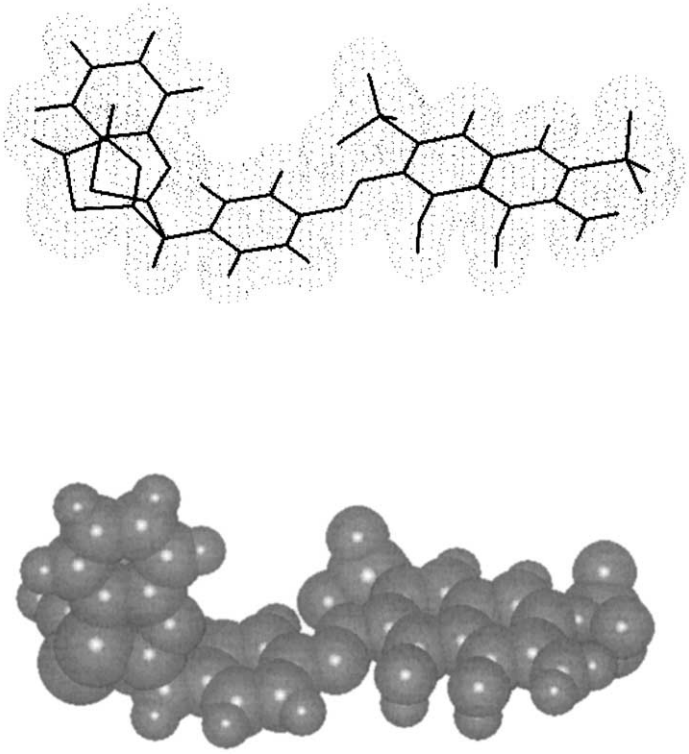


Fig. 4. Three-dimensional plot illustrating the superimposition of two selected dye molecules that contain both superimposible (right side) and nonsuperimposible moieties (left side).

and T3, respectively. Model quality decreases in the sequence T3 > T2 > T1. As alternative mapping focuses CoMSA analysis on a certain submotif, it accounts especially for electrostatic interactions of these motifs, while the rest of the molecule is ignored. Such an interpretation suggests that electrostatic interactions in the region of T3, which provides the best model of $q^2 = 0.981$, can explain a large share of the binding variance. This conclusion complies well with former CoMFA study [12] that indicates region T3 as the area of favorable interactions, while T1 was identified as the region of sterically unfavorable interactions. On the other hand, CoMFA analysis also suggested that electrostatic field is more important for binding than the steric one [5–12].

These results may be compared with our previous attempt to model the affinity of anionic azo and anthraquinone dyes [13]. It was observed in that study that dye–cellulose binding could be

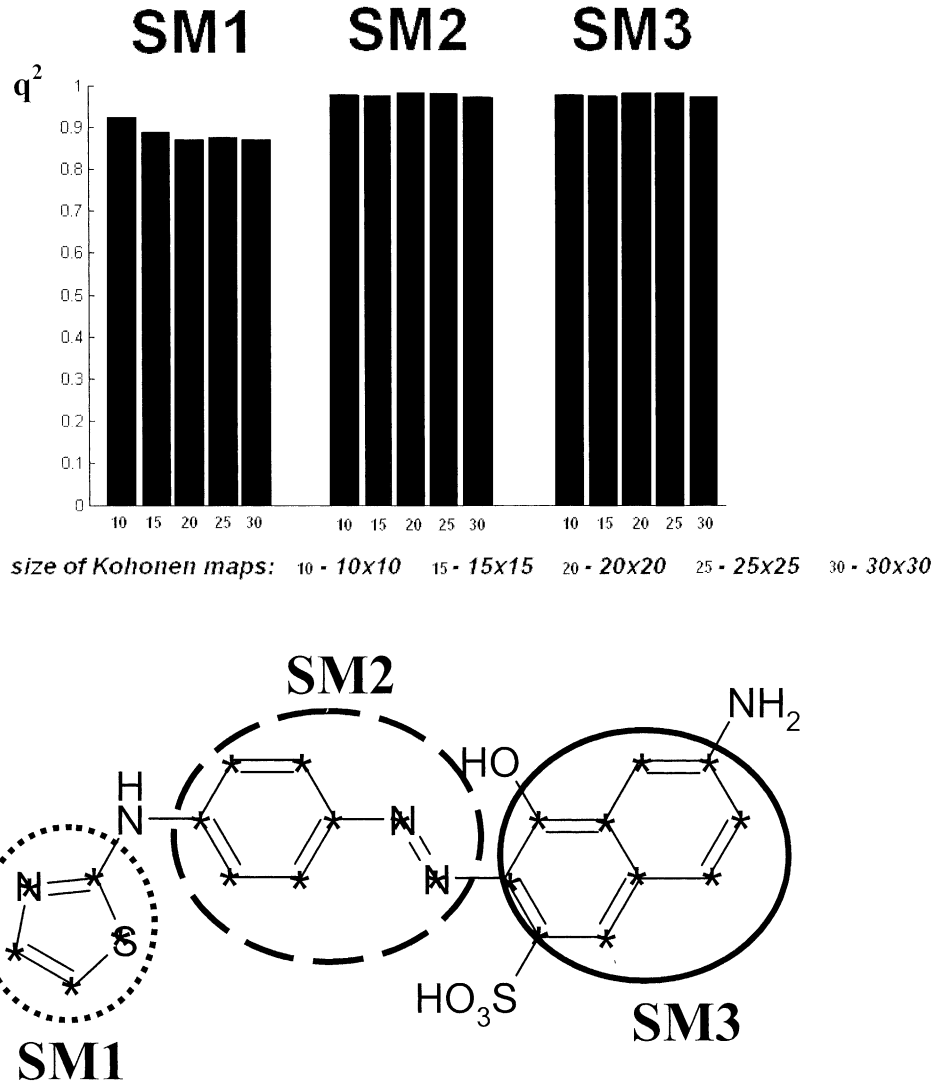



Fig. 5. The $q^2$ performances of CoMSA with different superimposition modes SM1–SM3 of the molecules. The circles indicate the molecule areas covered in the individual superimposition and the asterisks show individual atoms specified for covering.

modeled efficiently by CoMSA. Actually, the $q^2$ performances of these models do not depend upon the superposition mode and template selected. This was explained by a fact that relatively large cellulose cavities do not create steric hindrance to the dye molecules. In the current study, we can similarly observe that superposition mode only slightly influences model quality. Among three tested SMs only the SM1 decreases the $q^2$ performance. However, the results of alternative mapping depend upon the template selected, which makes a difference to the anionic azo and anthraquinone compounds [13]. Thus, the best $q^2$ performances for the CoMSA models based on T3 amount to $q^2 = 0.949$–$0.979$. The application of T2 slightly decreases these parameters to $q^2$ 0.828–0.897 and T3—results in significantly lower $q^2 = 0.582$–$0.646$ values. As alternative mapping simulates flexible binding, then such results show that heterocyclic monoazo dye–cellulose interactions

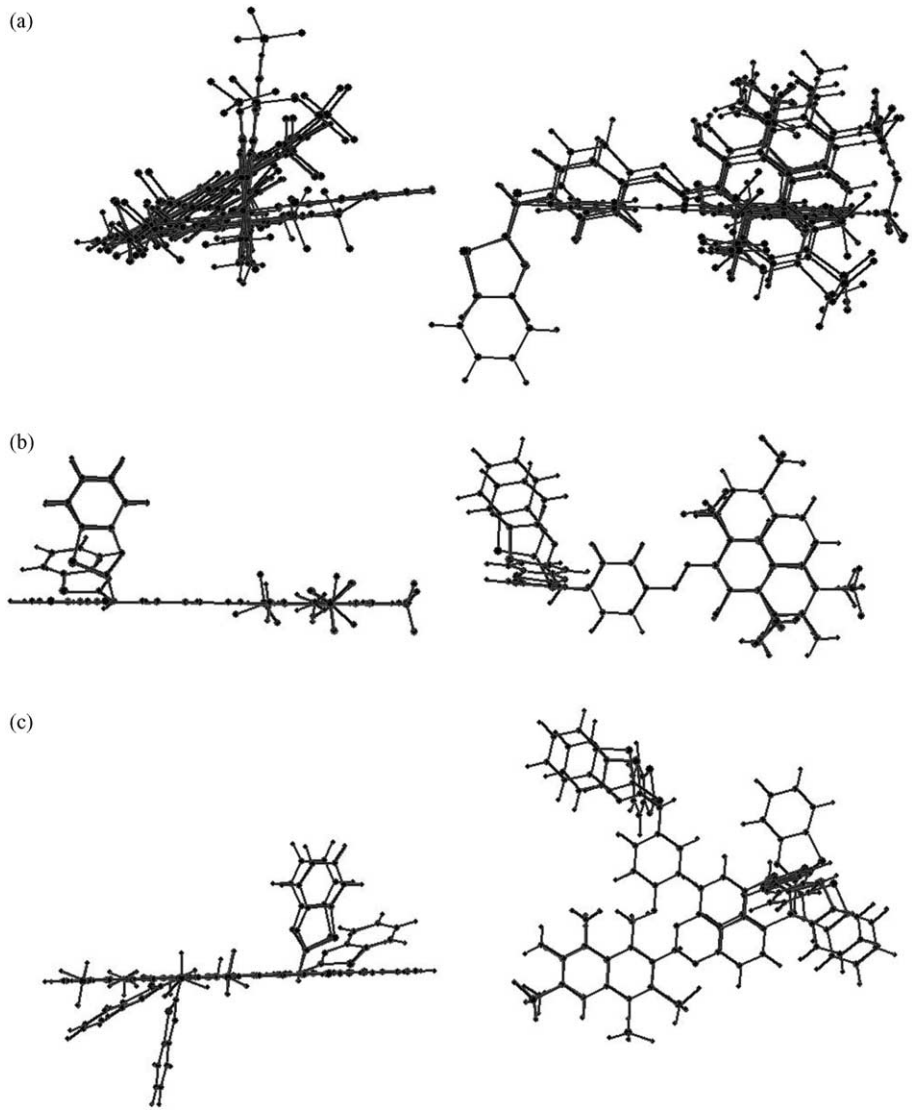
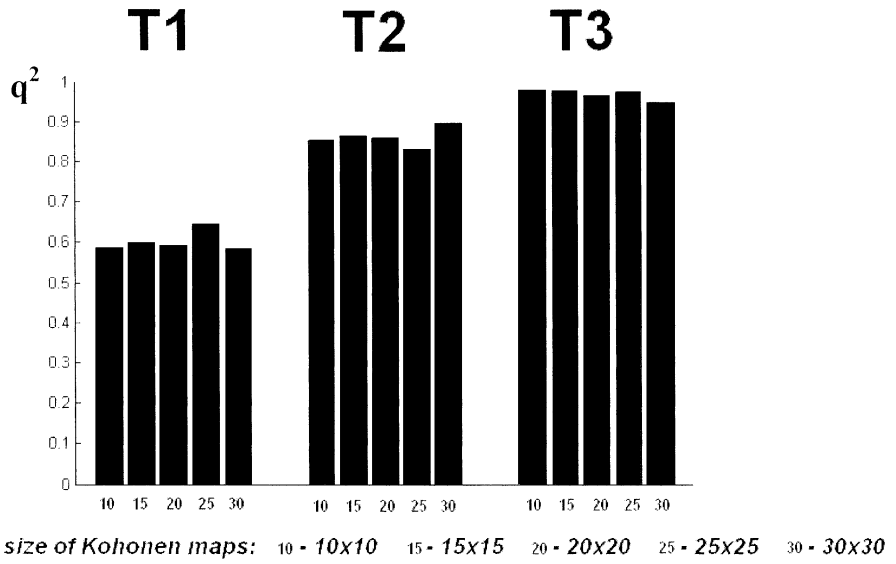



Fig. 6. Three-dimensional plots illustrating superimposition modes SM1–SM3.

do not involve a whole surface. This may suggest that steric requirements manifested by the cellulose receptor to monoazo heterocyclic dye series are more pronounced than in the case of anionic azo and anthraquinone compounds.

In Fig. 8, we show a bar plot illustrating the influence of the electrostatic matrix type used in CoMSA on the model quality. This effect differs with the template or superimposition mode applied. For the T3 template, i.e., the highly predictive models, the $q^2$ performances usually decreases in the sequence *en* = *nen* > *non–nen*. This can perhaps be explained by the T3 moiety fitting well into the cellulose "receptor". Thus, it is electrostatic interaction that determines dye affinity. This can be proved further by analyzing the SM3 superimposition mode performance bars. It is evident that the *en* matrices provide better models



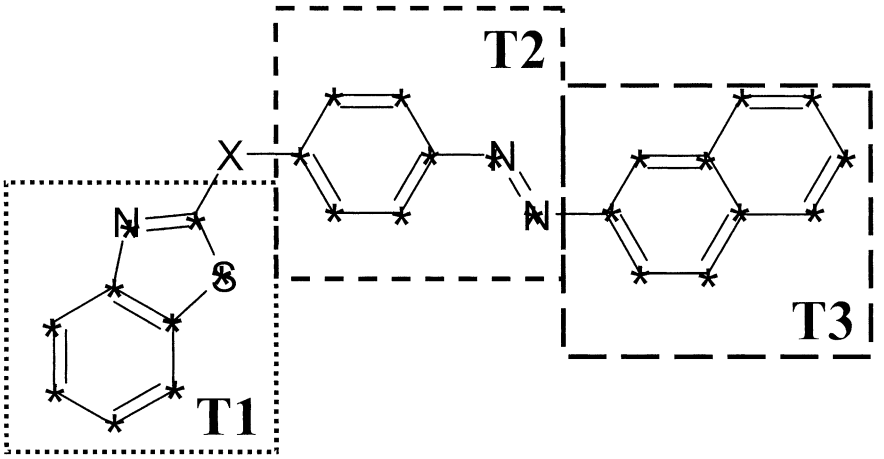




Fig. 7. The $q^2$ performances of CoMSA with different templates T1–T3 indicated on the two-dimensional molecular graph forming the basic skeleton of the series. The asterisks show individual atoms specified to be covered during CoMSA.

than others. The inclusion of shape effects in the model by the use of *non–nen* matrixes evidently decreases predictivity. To observe further the importance of the electrostatic interactions we modified the electrostatic matrices by changing all nonzero elements of the *en* matrices to 1 for positive values and −1 for negative values (*en*(1, 0,−1)), or to 1 for all non zero elements (*en*(1,0)). This significantly decreases model predictivities. Moreover, the *en*(1,0) matrices decreases this predictivity to the higher extend than *en*(1,0,−1). This seems to clearly indicate that electrostatic

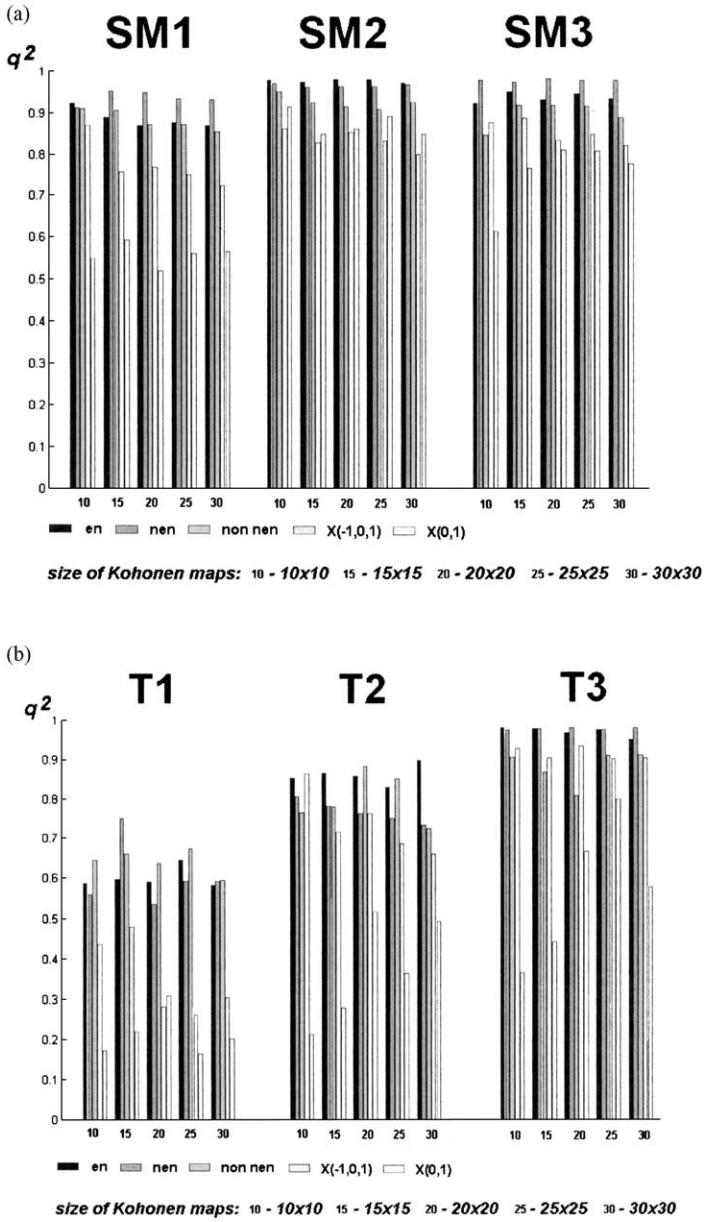



Fig. 8. The bar plots illustrating the $q^2$ performances of CoMSA with different electrostatic potential matrices using miscellaneous superposition modes and templates.

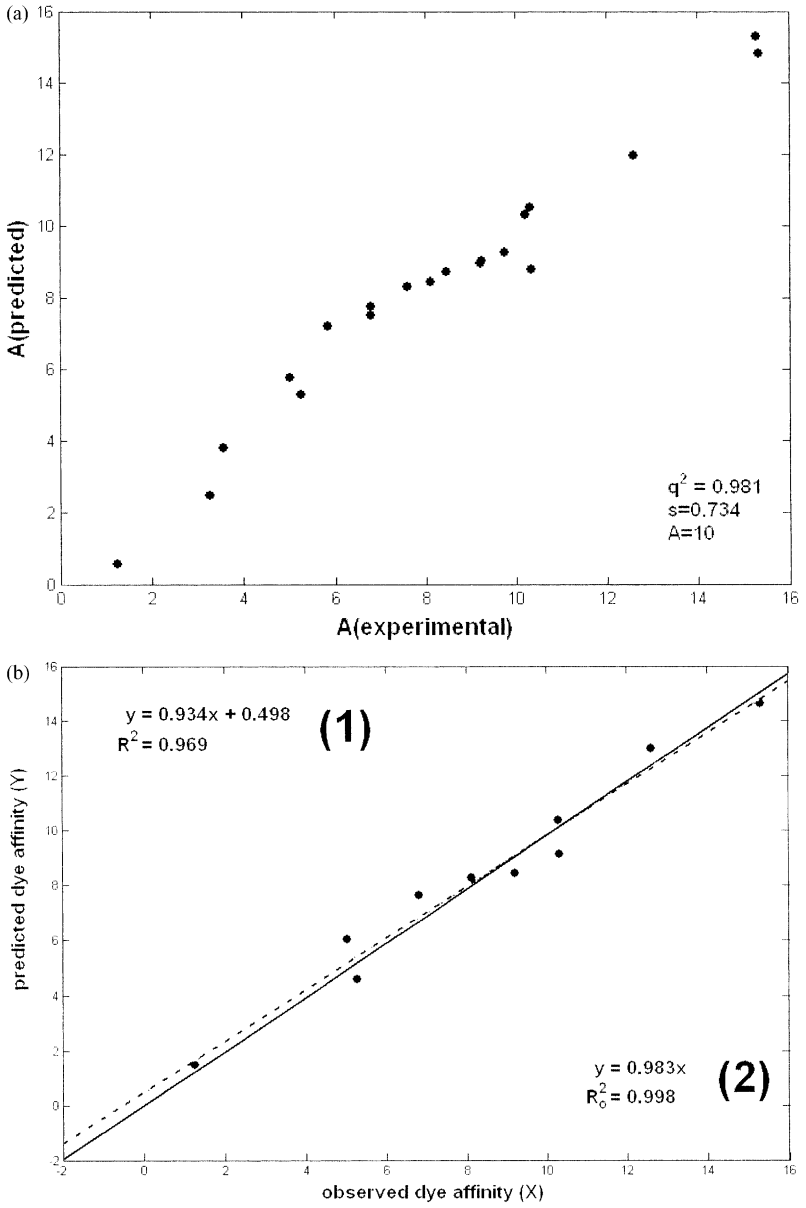



Fig. 9. The CoMSA model of dye affinity to cellulose, (a) Plot of experimental versus predicted (LOO-CV, with the optimal number (A = 10) of PLS components) dye affinities for heterocyclic monoazo dyes. (b) Validation of the best model by the Golbraikh–Tropsha criteria. We tried to keep the style of the presentation of the authors [40]. The regression between observed ($X$) and predicted ($Y$) activity values for the test set. The solid line shows the regressional equation given by (2). The dotted line illustrates the regression without the bias (1). The closer are these linear plots, the better is the model predictivity. Calculations after:

$$\mathrm{pred}_i^0 = k^*\mathrm{pred}_i; \qquad k = \frac{\sum \mathrm{obs}_i\mathrm{pred}_i}{\sum \mathrm{pred}_i^2}; \qquad R_0^2 = 1 - \frac{\sum\left(\mathrm{pred}_i - \mathrm{pred}_i^0\right)^2}{\sum(\mathrm{pred}_i - \mathrm{mean(pred)})^2}$$

where: upper index 0 relates to regression observed (Y) versus predicted ($X$); $k$—is a slope of the regression through the origin (2); and $R_0^2$—correlation coefficient for the regression of observed ($Y$) versus predicted ($X$) without *bias*. $[(R^2 - R_0^2)/R^2] < 0.1$ and $0.85 \leqslant k \leqslant 1.15$ as recommended by Golbraikh and Tropsha [40].

interactions are quite specific and can be precisely described in a 3D QSAR model.

In general, our results agree with the previous CoMFA studies that indicated polar attractions between dye molecule and cellulose as a decisive factor [12]. However, our results significantly outperform those obtained previously in CoMFA modeling. Thus, $q^2$ for CoMFA models ranges from 0.34 to 0.82 (cross-validated standard error $s = 3.37$–$1.84$), while for our best models $q^2$ reaches a value of $q^2 = 0.981$ and $s = 0.737$. Since the main difference between CoMFA and CoMSA strategies is that CoMSA, unlike the shapeless CoMFA approach concentrates in the areas defined by molecular surfaces, the better model performances and predictivities obtained in CoMSA seem to suggest that molecular surfaces comply much more accurately with the hypothetical interaction areas.

As current knowledge based on 3D QSAR modeling requires further verification by external prediction in column 7 of Table 1 we indicate the predicted values of dye–fiber affinity obtained in the LOO cross-validation procedure. We used the Galbraikh–Tropsha criterion [40] to verify the predictivity of the best models, which is shown in Fig. 9. This clearly indicates that CoMSA provides reliable and highly predictive models significantly outperforming those provided earlier by CoMFA method.

Recent literature on QSAR in dye chemistry has continued the debate about whether the pharmacophore concept can be used in this field. We hope that our work conclusively shows that dye *tinctophores* can contribute to the knowledge of the cellulose–dye interactions. Moreover, the high predictivity of the CoMSA models seems to indicate that these interactions are well defined and quite specific.

## 5. Conclusions

The application of the CoMSA method for modeling 3D QSAR of the cellulose affinity of the heterocyclic monoazo dyes allows us to obtain very predictive models significantly outperforming those reported previously in Comparative Molecular Field Analysis. We also used this method for the estimation of the relative importance of the steric and electrostatic interactions, which proved previous hypothesis that polar interactions are decisive for the dye–fiber affinity. Highly predictive models that were obtained indicate that dye–cellulose interactions are well-defined resembling drug–receptor interactions. This also means that a pharamacophore (or more precisely *tinctophore)* concept can be used efficiently for the description of the dye–fiber interactions.

## Acknowledgements

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussion and for facilitating access to the programs of CORINA, PETRA, SURFACE, and KMAP. The financial support of the KBN Warsaw: grants no: T08E02820, 4 T09A03424 and PBZ KBN-040 P04/08 is gratefully acknowledged.

## References

[1] Peters RH. Textile chemistry. In: The physical chemistry of dyeing. Vol. III. Amsterdam: Elsevier, 1975.
[2] Timofei S, Schmidt W, Kurunczi L, Simon Z. Dyes and Pigments 2000;47:5–16.
[3] French AD, Battista OA, Cuculo JA, Gray DG. In Kirk–Othmer encyclopedia of chemical technology. 4th ed. Vol. 5. New York: Wiley; 1993. p. 476.
[4] Timofei S, Schmidt W, Kurunczi L, Simmon Z, Sallo A. Dyes and Pigments 1994;24:267–79.
[5] Timofei S, Kurunczi L, Schmidt W, Fabian WMF, Simon Z. Quant Struct Act Relat 1995;14:444–9.
[6] Timofei S, Kurunczi L, Schmidt W, Simon Z. Dyes and Pigments 1995;29:251–8.
[7] Timofei S, Kurunczi L, Schmidt W, Simon Z. Dyes and Pigments 1996;32:25–42.
[8] Fabian WMF, Timofei S, Kurunczi L. J Mol Struct (THEOCHEM) 1995;340:73–81.
[9] Fabian WMF, Timofei S. J Mol Struct (THEOCHEM) 1996;362:155–62.
[10] Oprea TI, Kurunczi L, Timofei S. Dyes and Pigments 1997;33:41–64.
[11] Funar-Timofei S, Schüümann G. J Chem Inf Comput Sci 2002;42:788–95.
[12] Timofei S, Fabian WMF. J Chem Inf Comput Sci 1998; 38:1218–22.
[13] Polanski J, Gieleciak R, Wyszomirski M. J Chem Inf Comput Sci 2003;43:1754–62.

[14] Polanski J. Acta Pol Pharm 1999;56:80–4.
[15] Polanski J, Walczak B. Comput and Chem 2000;24:615–25.
[16] Polanski J, Gieleciak R, Bąk A. J Chem Inf Comput Sci 2002;42:184–91.
[17] Polanski J, Gieleciak R. J Chem Inf Comput Sci 2003; 43:656–66.
[18] Alberti G, Seu G. Ann Chim (Rome) 1983;73:737–40.
[19] Alberti G, Cerniani A, De Giorgi M, Seu G. Ann Chim (Rome) 1983;73:265–72.
[20] Alberti G, Cerniani A, De Giorgi M, Seu G. Tinctoria 1980;5:141–5.
[21] Kohonen T. Self-organization and associative memory. 3rd ed. Berlin: Springer; 1989.
[22] Gerstein M, Greenbaum D, Luscombe MN. Method Inform Med 2001;40:346–58.
[23] Brazma A, Vilo J. FEBS Letters 2000;480:17–24.
[24] Toronen P, Kolehmainen M, Wong G, Castren E. FEBS Letters 1999;451:142–6.
[25] Anzali S, Gasteiger J, Holzgrabe U, Polanski J, Sadowski J, Teckentrup A, et al. Perspect Drug Discov Design 1998; 11:273–99.
[26] Zupan J, Gasteiger J. Neural networks and drug design for chemists. 2nd ed. Weinheim: VCH; 1999.
[27] Gasteiger J, Li X, Rudolph Ch, Sadowski J, Zupan J. J Am Chem Soc 1994;116:4608–20.
[28] Polanski J. J Chem Inf Comput Sci 1997;37:553–61.
[29] Anzali S, Barnickel G, Krug M, Sadowski J, Wagener M, Gasteiger J, et al. J Comp-Aided Mol Design 1996;10:521–40.
[30] Polanski J. Adv Drug Deliv Rev 2003;55:1149–62.
[31] Polanski J, Gasteiger J, Jarzembek K. Comb Chem and HTS 2000;3:481–95.
[32] Hasegawa K, Matsuoka S, Arakawa M, Funatsu K. Comput and Chem 2002;26:583–9.
[33] Gasteiger J. CORINA for the information. Available from: http://www.mol-net.de.
[34] Sadowski J, Gasteiger J. Chem Reviews 1993;93:2567–81.
[35] Sadowski J, Gasteiger J, Klebe G. J Chem Inf Comput Sci 1994;34:1000–8.
[36] Gastaiger J, Sailer H. Angew Chem 1985;97:699–701.
[37] Gasteiger J, Marsili M. Tetrahedron 1980;36:3219–28.
[38] Gasteiger J. Match3D; KMAP for the information. Available from: http://www2.ccc.uni-erlangen.de.
[39] Polanski J, Zouhiri F, Jeanson L, Desmaële D, d'Angelo J, Mouscadet JF, et al. J Med Chem 2002;45:4647–54.
[40] Golbraikh A, Thropsha A. J Mol Graph Mod 2002; 20:269–76.