ナフトール-AS誘導体のMLRとANNによる構造毒性相関

Structure-Toxicity Relationship Study of Some Naphthol-AS Derivatives by MLR and ANN

Takahiro SUZUKI[1], Simona FUNAR-TIMOFEI[2],
Sergiu Adrian CHICU[3], Georgeta-Maria SIMU[4], Luminita CRISAN[2]

Abstract

The structure-toxicity relationships for a series of 22 arylamides of Naphthol – AS type are presented by computational approaches. The toxicity of the model compounds was determined by using the *Hydractinia echinata* (hydrozoa) test system. Measured toxicity values, Mlog(1/MRC_{50}), were considered as dependent variable and were related to structural features obtained by molecular and quantum mechanics calculations by multiple linear regression (MLR) and artificial neural network (ANN) approaches. Variable selection was carried out by genetic algorithm. The obtained models showed that arylamide toxicity was influenced by molecular geometry, as well as by compound polarity.

Keywords: *Hydractinia echinata*, Naphthol-AS, structure-toxicity relationship, GA-MLR, ANN

1. Introduction

The anilide of the 2-hydroxy-3-naphthoic acid, also known as Naphthol-AS, exhibits an exceptional substantivity for cotton fibres. Due to this feature, this compound is extensively used as coupling component in the synthesis of some disazo compounds which can be obtained directly on the fibre, at low temperatures. These dyes are essentially insoluble and are ranged in the class of vat dyes or azotols. They exhibit nice and bright shades and are characterized by good fastness to light and washing, as compared to those presented by the anthraquinonic dyes.

From structural point of view, the Naphthols-AS exhibit most of the features pointed-out by Elkins (Elkins et al., 2000), conditions which are necessary to the binding of these

[1] Natural Science Laboratory, Toyo University, 5-28-20 Hakusan, Bunkyo-ku, Tokyo 112-8606, Japan
[2] Institute of Chemistry Timisoara of the Romanian Academy, B-dul Mihai Viteazul 24, RO-300223 Timişoara, Romania
[3] Siegstr. 4, 50859 Köln/Deutschland tel.: 02234/78530 e-mail: lchi@gmx.de
[4] University of Medicine and Pharmacy Victor Babeş Timişoara, Faculty of Pharmacy, Piata Eftimie Murgu 2-4, RO-300034 Timişoara, Romania

compounds to a peptidase: one carbonylic oxygen atom, a hydrophobic center, and an aminic nitrogen atom (hydrogen acceptor area). A low energetic conformation and an adequate solubility are also required (Milne et al., 1998).

Due to the presence of the strong polarisable carbonylic bond, the hydrolysis reaction of the cromophoric group (amidic) occurs perhaps by *Hydractinia echinata* (hydrozoa) enzymatically, by a mechanism similar to the hydrolysis of the peptidic groups (Alberts et al., 1983). Besides, the rate of the partial biodegradation products formation will depend on the rate and intensity of the substrate-enzyme interaction.

The aim of the present work was to relate the experimental Naphthol-AS toxicity values obtained by the *H.echinata* test system to structural descriptors derived from minimum energy structures obtained by molecular and quantum mechanics calculations, in order to have an insight on structural features that influence the toxicity. Correlations with toxicity values were carried out by multiple linear regression, combined with genetic algorithm (GA-MLR), which was used for the variable selection. Artificial neural networks (ANN) using the same set of descriptors were also applied to the structure-toxicity problem.

2. Materials and Methods

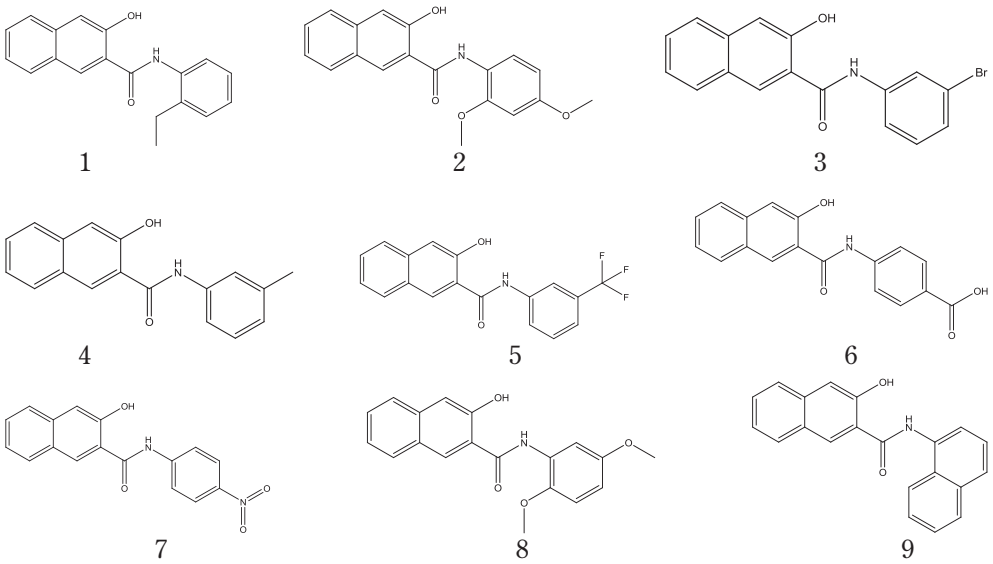

Fig 1. Arylanides of Naphthol – AS type structure

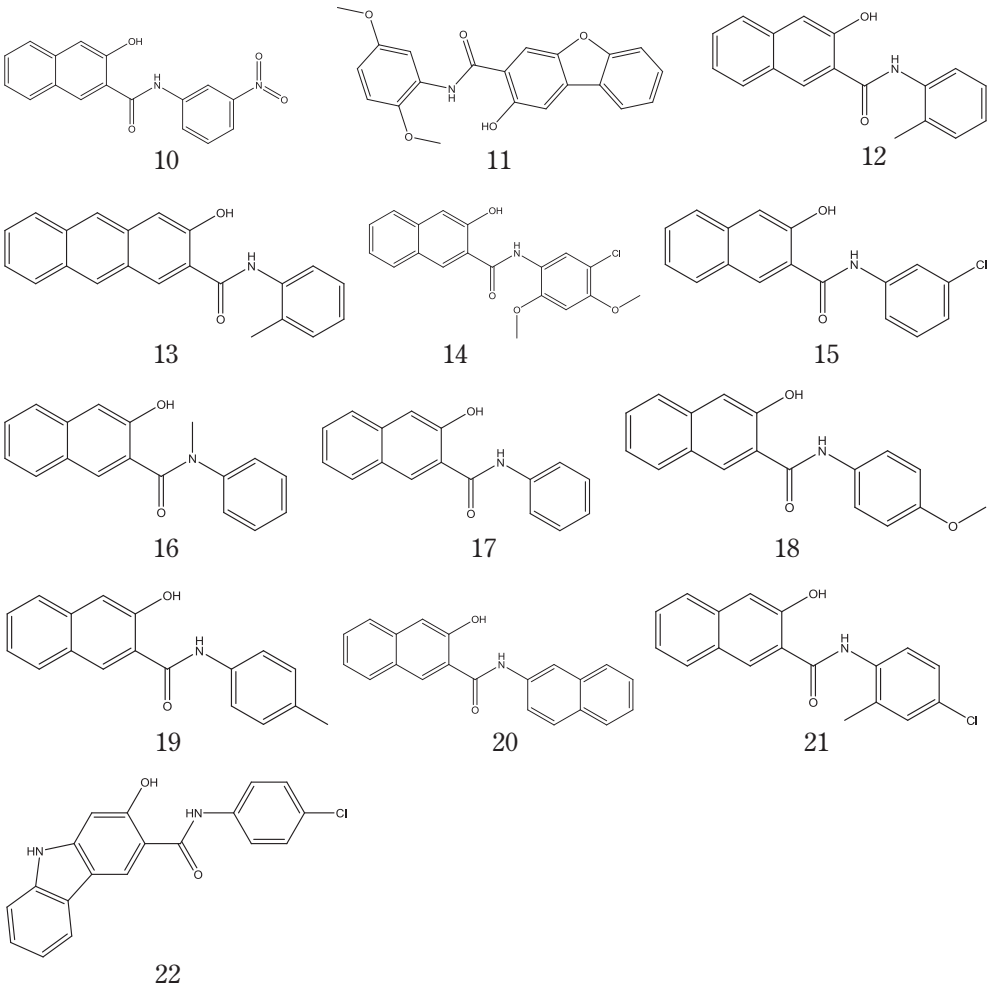

Fig 1. (continued)

The test substances, 22 Naphthol-AS derivatives (Fig. 1), were synthesized at the Institute of Chemistry of Timisoara of the Romanian Academy, Romania (Chicu et al., 2011). They were used such as or as solutions of known concentrations in methanol or synthetic seawater.

The test conditions and the test method were identical to those described in previous works (Chicu et al., 2000; Chicu et al., 2009; Chicu and Simu, 2009). The toxicity of the test substances represents the concentration MRC_{50} (Metamorphosis Reducing Concentration, in $mol \cdot L^{-1}$), expressed as the reciprocal value of its logarithm. $Mlog(1/MRC_{50})$ at which the frequency of metamorphosis induction of *H. echinata* from the larvae to polyp, was reduced by 50 % in comparison with our test control and was used as dependent variable.

3. QSAR modeling

Structural parameters

The molecular structures of 22 arylamides of Naphthol – AS type (presented in Fig. 1) were modeled previously (Chicu et al., 2011) by the conformational search ability of the Omega v.2.3.2 (OpenEye Scientific Software, Santa Fe, NM 87507) program (only structures having toxic effect on *H. echinata* test system were considered).

From the conformational search of each molecule the minimum energy structure was used to derive structural descriptors. Thus, twenty-two types of descriptors were calculated by the Dragon software, e.g. Dragon Professional 5.5/2007, Talete S.R.L., Milano, Italy, like: constitutional, functional groups counts, topological descriptors, Burden eigenvalues, eigenvalue-based indices, Galvez descriptors (topological charge indicies), Getaway descriptors: R8u+ - R maximal autocorrelation of lag 8 / unweighted, ISH - standardized information content on the leverage equality Randic descriptors (Randic molecular profiles), RDF descriptors (radial distribution function descriptors): RDF075m - Radial Distribution Function - 7.5 / weighted by atomic masses, RDF080u - Radial Distribution Function - 8.0 / unweighted, RDF055p - Radial Distribution Function - 5.5 / weighted by atomic polarizabilities, RDF060v - Radial Distribution Function - 6.0 / weighted by atomic van der Waals volumes, RDF040p - Radial Distribution Function - 4.0 / weighted by atomic polarizabilities; MWC (Molecular walk counts path counts – atomic and molecular descriptors) and 3D-MoRSE (3D-molecule representation of structure based on electron diffraction descriptors), atom-centred fragments, information indices, edge adjacency indices, topological charge indices, connectivity indices, 2D-autocorrelations, molecular properties, 2D binary fingerprints, and 2D frequency fingerprints: F04[C-N] - frequency of C-N at topological distance 4, F08[O-O] – frequency of O-O at topological distance 8.

In addition, hydrofobicities were computed by the ALOGPS 2.1 program (Tetko and Tanchuk, 2002): logKow (Meylan and Howard, 1995), AlogP, ClogP (see: ClogP v.4.0, Biobyte, Claremont, CA, USA, available on http://146.107.217.178/lab/alogps/start.html), IA _logP and IA_logS – octanol/water partition coefficient, respectively the water solubility calculated by Interactive Analysis, using neural networks technology http://146.107.217.178/lab/alogps/start.html), XlogP (see: http://146.107.217.178/lab/alogps/start.html), COSMO frag (Hornig and Klamt, 2005).

Single point quantum chemical calculations were applied by the MOPAC 2009 software (MOPAC2009, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, http://OpenMOPAC.net, 2008) to the conformations of minimum energy. Several descriptors derived from MOPAC calculations were used, like: dipole moment, HOMO and LUMO energies calculated for the entire molecule and for the hydroxilic ($LUMO_{hydroxyO}$),

amidic oxygen and nitrogen amidic atoms, electrophilicity index, COSMO area and volume (Klamt and Schüümann 1993), Mulliken electronegativity, absolute hardness, nucleophilic (Dn(r)) and electrophilic (De(r)) delocalisabilities, atom polarisability (piS(r)) calculated for the entire molecule and for the hydroxilic and amidic oxygen and nitrogen amidic atoms and charge for the hydroxilic ($q_{hydroxyO}$), amidic oxygen, and nitrogen amidic atoms, and hydroxilic and amidic hydrogen atoms.

Multiple linear regression (MLR)

Multiple linear regression relates one experimental variable y_k to one or several structural variables x_i by the equation (Wold and Dunn III, 1983):

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

where b represents regression coefficients and e the deviations and residuals. MLR calculations were performed by the STATISTICA 7.1, Tulsa, StatSoft Inc, OK, USA and MobyDigs programs (Todeschini et al., 2004a).

Genetic algorithm (GA) (Rogers and Hopfinger 1994) was performed to search the feature space and select descriptors relevant to toxicity. The first step of GA is to generate a set of solutions (chromosomes) randomly which is called an initial population. Then a fitness function is used to evaluate the fitness of these individuals, and a new population is formed consisting of the fittest chromosomes as well as offspring of these chromosomes based on the notion of survival of the fittest. The Kubinyi fitness function (Kubinyi, 1994) was used in our study as the RQK fitness function (Todeschini et al., 2004b). Then crossover and mutation operations are performed to generate new individuals. In the subsequent selection stage, the fittest individuals evolve to the next generation. These steps of evolution continue until the stopping conditions are satisfied. In the current work, the models were built using the simple MLR method with the selected variables from GA, called GA-MLR.

Artificial neural networks (ANNs)

The artificial neural networks have an inherent ability to provide non-linear and cross product terms for QSAR modeling. The ANNs are especially useful when a rigid theoretical basis and/or mathematical relationship to describe a phenomenon to be modeled are not available in advance.

The three-layer ANNs with the back-propagation of errors were employed in this study. Since the theory and practical application of the ANN are popular, an explanation of the methodology can be delegated to the literature (Zupan and Gasteiger, 1999). The most commoly used log sigmoid transfer function and the delta rule for the error correction formula were used in the networks. The ANN calculations were carried out by using our inhouse program.

Model validity

All the statistical tests were performed at a significance level of 5 % or less. In GA-MLR calculations outliers were tested by estimating the standardized residuals of less than -3.0 or more than +3.0 (Frank and Althoen, 1995) and by the value of residual greater than three times the value of standard error in calculation (Todeschini and Consonni, 2000a), as implemented in the MobyDigs program (Todeschini and Pavan, 2004a). The Kubinyi fitness function (Kubinyi, 1994) was, also, used to check the goodness of fit of the obtained GA-MLR models.

The goodness of prediction of the GA-MLR models was checked by the Akaike Information Criterion (AIC) (Gentleman and Wilk, 1975), the multivariate K correlation index (Todeschini and al, 1999), Y-scrambling (Lindgren and al, 1996), external validation (Todeschini and Consonni, 2000b). Y scrambling was applied to exclude the possibility of chance correlation and to check for reliability and robustness by permutation testing: new models were recalculated for randomly reordered responses (Y scrambling). The resulting models obtained with randomized responses should have significantly lower q^2 values than the proposed ones because the relationship between the structure and response is broken. Y scrambling was performed by response scrambling with maximum iterations of 500, and then the mean values of $R^2_{Yscrambling}$ ($a(r^2)$) and $Q^2_{Yscrambling}$ ($a(q^2)$) were reported. The predictive power of a QSAR model can be estimated by the external q^2_{est} defined as follows:

$$q^2_{ext} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y}_{tr})^2} \qquad (2)$$

where y_i and $\hat{y}_i$ are the experimental and predicted values of the dependent variable for the test set, $\bar{y}_{tr}$ is the mean value of the dependent variable for the training set, and m is the compounds number of the test set.

All these calculations were performed by the MobyDigs software. The leave-one-out cross-validation procedure (Wold, 1978) was also employed to check the robustness of the model.

To avoid models with collinearity without prediction power, regression models were calculated only for variable subsets with an acceptable multivariate correlation applying the recently proposed QUIK (Q Under Influence of K) procedure based on the K multivariate correlation index (Todeschini et al., 1999). Only models with a global correlation of [XY] block (K_{XY}) greater than the global correlation of the X block (K_{XX}) variable can be accepted, where X is the descriptor matrix and Y is the dependent variable. To each model, the K_{XY} and K_{XX} values were calculated. Several commonly used statistic terms were adopted to check the reliability, robustness and stability of the proposed model such as correlation coefficient (r^2), leave-one-out (LOO) cross-validated q^2_{LOO}, root mean squared error for the training set (SDEC) and predictive set (SDEP) (Consonni and al, 2009).

Tools of regression diagnostic as residual plots and Williams plots were used to check the quality of the best models and define their applicability domain using the Mobydigs software. Residual plot shows validated residuals versus response values and enables the search for outliers and to verify the assumption of the GA-MLR method on the normal error distribution, therefore this plot is a tool to evaluate the existence of a linear relation between variables and response. Leverages of test compounds were calculated to check their distance from the model experimental space; the greater the distance the more unreliable the predicted response (Frank and Todeschini, 1994).

4. Results and Discussion

GA-MLR results

The principles for assessing the validity of QSARs for regulatory purposes, e.g. the OECD Principles for the Validation of (Q)SARs, http://www.oecd.org/dataoecd/33/37/37849783.pdf (last accessed November 2008) were verified. Structural descriptors were derived from the arylamide structures of minimum energy obtained by conformational analysis. A training set of 17 compounds: no. 1, 3, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22 (Fig.1) and a test set of 5 compounds: no.: 2, 4, 5, 7, 21 were ordered by their decreasing toxicity values. The test set was chosen according to a random selection through toxicity sampling.

Starting from the total set of calculated descriptors, GA-MLR analysis has been applied to model the toxicity of the arylamides. Intercorrelated variables having correlation coefficient r = 0.7 were removed. Variable selection was carried out to non-correlated variables by the genetic algorithm included in the MobyDigs program (Todeschini and al, 2004a), using the RQK fitness function (Todeschini and al, 2004b), with Kubinyi fitness function as constrained function to be optimised, a crossover/mutation trade-off parameter T = 0.5 and a model population size P = 50 and maximum generations of 2000. Compound 22 was found as an outlier.

Standard deviation error in prediction (SDEP), standard deviation error in calculation (SDEC), standard error of estimate (SE), inter-correlation of selected descriptors (K_{XX}) and the correlation of the X block with response (K_{XY}) are also reported for each model, together with the coefficient of determination (r^2), cross-validated explained variance by leave-one-out (q^2_{LOO}) and external validation (q^2_{est}).

The presence of outliers (i.e. compounds with crossvalidated standardized residuals greater than three standard deviation units), and chemicals very influential in determining model parameters were verified by the Williams plot (Mobydigs). The leverage approach was also applied for the definition of the chemical domain of each model (Tropsha et al., 2003). The GA-MLR models are presented in Table 1. Compound 22 was found as an outlier.

Table 1. Models obtained by GA-MLR*

Model	Descriptors	r^2	q^2_{LOO}	q^2_{boot}	q^2_{est}	$a(r^2)$	$a(q^2)$	r^2_{adj}	SDEP	SDEC	F	SE	AIC	Kx	Kxy	FIT
1	RDF075m H1e B08[O-O]	0.718	0.610	0.477	0.645	0.488	0.194	0.649	0.604	0.513	10.23	0.592	0.585	41.45	46.58	1.223
2	RDF075m ISH F04[C-N]	0.770	0.644	0.513	0.758	0.132	-0.559	0.717	0.666	0.535	14.53	0.612	0.605	28.18	47.37	1.679
3	ISH R1e+ Mor28p	0.753	0.592	0.451	0.360	0.148	-0.651	0.691	0.618	0.481	12.17	0.555	0.514	19.02	36.26	1.455

* r^2 – correlation coefficient, q^2_{LOO} – leave-one-out crossvalidation parameter, q^2_{boot} – bootstrapping parameter, q^2_{est} – external q^2, $a(r^2)$ and $a(q^2)$ – Y-scrambling variables, r^2_{adj} –adjusted r^2, SDEP – standard deviation error in prediction, SDEC – standard deviation error in calculation, F- Fischer test, SE – standard error of estimate, AIC- Akaike Information Criterion, the multivariate K correlation indices (Kx-the multivariate correlation index of the matrix of X descriptors and Kxy - the multivariate correlation index of the matrix of X descriptors and Y response variable), FIT- the Kubinyi fitness function; results presented in parenthesis correspond to models from which the outlier wase removed.

From all these models, following model was chosen with best results for the 16 compounds in the training set, after removing the outlier:

$$\begin{aligned} &M\log(1/MCR_{50}) = -8.73(\pm 2.88) - 0.20(\pm 0.08)RDF075m + 6.19(\pm 1.31)H1e \\ &-1.08(\pm 0.45)B08[O-O] \\ &n = 16; r^2 = 0.718; q^2_{LOO} = 0.610; q^2_{boot} = 0.477; q^2_{ext} = 0.645; r^2_{adj} = 0.649; a(r^2) = 0.488; \\ &q(r^2) = 0.194; SDEC = 0.513; SDEP = 0.604; K_X = 41.45; K_{XY} = 46.58; F(3,12) = 10.23; \\ &SE = 0.592; AIC = 0.585; SDEP(ext) = 0.576; RMS = 0.597 \end{aligned} \quad (3)$$

After removing the outlier, the applicability domain of the model with 16 compounds was evaluated by leverage analysis expressed as Williams plot (see Fig. 2), in which the standardized residuals and the leverage values were plotted. This plot confirms the absence of outliers and influential points (the leverage average value being of 0.250).

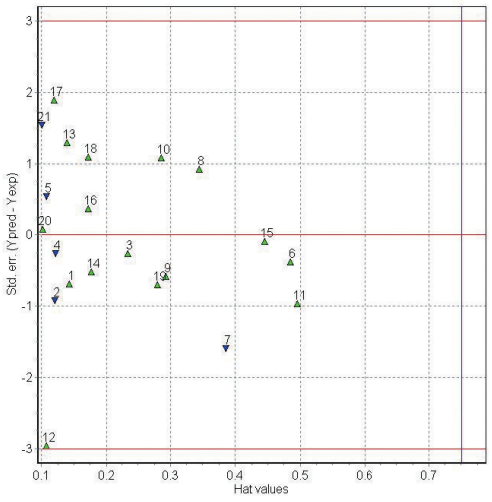

Fig. 2. Williams plot: jackknifed residuals of the best model (eq. 3) versus leverages. Training compounds are marked by green triangles and test compounds by blue triangles

Predicted Mlog(1/MRC_{50}) values of external validation compounds are reported in Figure 3 together with predicted Mlog(1/MRC_{50}) values of the training compounds. The response plot confirms the model quality, because it shows a good alignment of the heterogeneous arylamides along the optimal line.

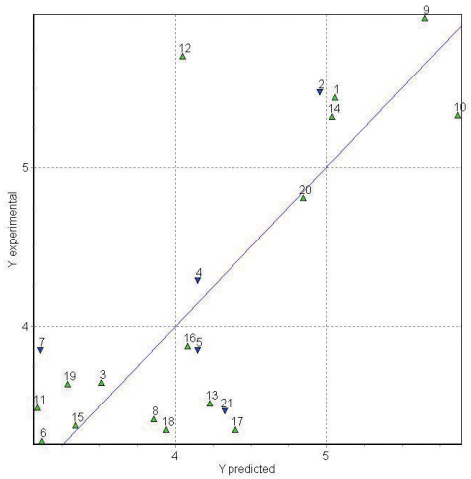

Fig. 3. Experimental versus predicted Mlog(1/MRC_{50}) values. Training compounds are marked by green triangles and test compounds by blue triangles

The predictive power of the model is guaranteed by the validation made by the leave-more-out procedure; a q^2_{LOO} value of 61.02 % was obtained, highlighting the stability of the model. The quality of the model was also tested on an external validation set of 5 compounds belonging to the model applicability domain mainly on the basis of their leverages. The q^2_{ext} value of 64.5 %, suggests that the efective prediction power of the model is slightly better than that obtained by internal validation.

The selection criterion used in this study is that the model should have higher cross-validated q^2_{LOO}, higher external predictive ability, least difference between internal and external predictive ability, the fewer chemicals outside the chemical domain and the fewer chemicals with large relative errors. On the basis of the above principles, eq. 3 was selected as the best single model. From all the statistical parameters, it can be seen that the proposed model is stable, robust and predictive.

The studied arylamides have similar structures. Looking at the applicability domain of the found model compounds 22 (in which the hydrogen bond formed beween the hydrogen hydroxy group of the naphthalene moiety and the carbonyl oxygen atom is absent) was found as outlier, according to the standardised residuals.

The set of molecular descriptors in the above MLR model was used to develop the nonlinear models by ANNs based on the same set of 16 training and five test compounds. The performance of the back-propagation ANNs, the number of hidden neurons and the

number of iteration cycle are the two significant parameters influencing the performance of this model. The optimum number of neurons in the hidden-layer and the epochs were selected as 3 and 45000, respectively, with learnig rates and gain set of 0.5 and 2, respectively. The best network gave a RMS error value of 0.456 for the training set and 0.494 for the prediction set, respectively, indicating somewhat better statistical results in comparison to the above presented MLR model. However, the ANN model's predictive power was inferior to that of the MLR model.

Arylamide toxicity is influenced by conformational changes and by the bond lengths that account for atom types and bond multiplicity, information, about bond distances, ring types, planar and non-planar systems and atom types. For instance increased values of RDF075m RDF descriptor, which encodes the Radial Distribution Function - 7.5 / weighted by atomic masses and of B08[O-O] 2D binary fingerprint descriptor (which defines the presence/ absence of O-O at topological distance 08) yield more toxic compounds. The H1e Getaway descriptor, which represents the H autocorrelation of lag1/weighted by atomic Sanderson electronegativities and encodes both the geometrical information given by the molecular influence matrix H and the topological information given by the molecular graph, weighted by selected atomic weights, decrease the arylamide toxicity.

Genetic algorithm was applied for the selection of parameters relevant to the compound toxicity and GA-MLR models were selected according to the OECD principles for assessing the validity of QSARs for regulatory purposes. Models with satisfactory robustness were obtained. Structural parameters which describe molecular conformational changes (including information about bond distances, ring types, planar and non-planar systems and atom types) and polarity influence the arylamide toxicity.

4. Conclusion

Multiple linear regressions combined with genetic algorithm for variable selection was used to correlate the toxicity of the model compounds, determined by using the *Hydractinia echinata* (hydrozoa) test system, with structural features of Naphthol-AS derivatives. Following descriptors: 2D binary fingerprint, RDF and Getaway descriptors were present in the final MLR model with acceptable statistical results. The MLR model was then compared to those obtained by ANN. The most stable models obtained by both approaches had comparable statistical results.

Acknowledgements

This project was financially supported by Project 1.1 of the Institute of Chemistry Timisoara of the Romanian Academy.

References

Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J.D. (1983) *Molecular Biology of the Cell*. (Garland Publishing, Inc.), pp. 131.

Chicu, S.A; Herrmann, K.; Berking, S. (2000). An approach to calculate the toxicity of simple organic molecules on the basis of QSAR analysis in *Hydractinia echinata* (Hydrozoa, Cnidaria). *QSAR* **19**, 227-236.

Chicu, S. A.; Grozav, M.; Kurunczi, L.; Crisan, M. (2009). SAR for amine salts of carboxylic acids to *Hydractinia echinata* test system. *Rev Chim -Bucharest* **5**, 582-587.

Chicu, S. A.; Simu, G. M. (2009). *Hydractinia echinata* test system. I. Toxicity determination of some benzoic, biphenylic and naphthalenic phenols. Comparative SAR-QSAR study. *Rev Roum Chim* **54**(8), 659-669.

Chicu, S.A.; Funar-Timofei, S.; Simu, G. M. (2011) *Hydractinia echinata* test system. II. SAR toxicity study of some anilide derivatives of Naphthol-AS type, *Chemosphere*, **82**, 1578–1582.

Consonni V. Ballabio D. Todeschini R. (2009). Comments on the definition of the *Q2* parameter for QSAR validation, *J Chem Inf Model* **49**, 1669–1678.

Elkins S.; Waller C.L.; Swaan P.W; Cruciani G.; Wrighton S.A; Wikel, J. H. (2000). Progress in predicting human ADME parameters in silico. *J Pharmacol Toxicol* **44**, 251-272.

Frank H.; Althoen S.C. (1995). *Outliers*. In: Statistics: Concepts and Applications, Cambridge, Great Britain, Cambridge University Press, pp. 142-143.

Frank I. E. Todeschini R. (1994). *The Data Analysis Handbook*. Elsevier, Amsterdam, The Netherlands.

Gentleman J.F.; Wilk M.B. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics* **31**, 387-410.

Hornig M.; Klamt A. (2005). COSMO frag: a novel tool for high-throughput ADME property prediction and similarity screening based on quantum chemistry. *J Chem Inf Comp Sci* **45**(5), 1169–1177.

Klamt A.; Schüümann G. (1993). COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J Chem Soc*: *Perk T* **2**, 799-805.

Kubinyi H. (1994). Variable selection in QSAR studies. I. An evolutionary algorithm. *Quant Struct-Act Rel* **13**, 285–294.

Lindgren F.; Hansen B.; Karcher W.; Sjöström M.; Eriksson L. (1996). Model validation by permutation tests: Applications to variable selection. *J Chemometr* **10**, 521-532.

Meylan W. M.; Howard P.H. (1995). Atom/fragment contribution method for estimating octanol-water partition coefficients. *J Pharm Sci* **84**, 83-92.

Milne G.W.A; Nicklaus M.C; Wang S. (1998). Pharmacophores in drug design and discovery. *SAR QSAR Environ Res* **9**, 23-38.

Rogers D.; Hopfinger A. J. J. (1994). Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property Relationships, *Chem Inf Comput Sci* **34**, 854-866.

Tetko I. V.; Tanchuk V. Y. (2002). Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 Program. *J Chem Inf Comp Sci* **42**, 1136-1145.

Todeschini R.; Consonni V.; Maiocchi A. (1999). The K correlation index theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* **46**, 13-29.

Todeschini R.; Consonni V., (2000). *Handbook of Molecular Descriptors*. Wiley, Weinheim, pp. 369a , 463b

Todeschini R.; Ballabio B.; Consonni V.; Mauri A.; Pavan M. (2004a) MOBYDIGS – Version 1 – Copyright (c) 2004 by TALETE srl, Milano, Italy

Todeschini R.; Consonni V. ; Mauri A. ; Pavan M. (2004b). Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal Chim Acta* **515**, 199-208.

Tropsha A.; Gramatica P.; Gombar V. K. (2003). The importance of being Earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci* **22**, 69–76.

Wold S. (1978). Cross validatory estimation of the number of components in factor and principal components models, *Technometrics* **20**, 397-405.

Wold S.; Dunn III W.J. (1983). Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. *J Chem Inf Comp Sci* **23**, 6-13.

Zupan, J.; Gasteiger, J. (1999). *Neural Networks for Chemistry and Drug Design*, 2nd Edition, Wiley-VCH, Weinheim.

要　旨

ナフトール -AS 誘導体の MLR と ANN による構造毒性相関

鈴木孝弘 , Simona FUNAR-TIMOFEI, Sergiu Adrian CHICU, Georgeta-Maria SIMU, Luminita CRISAN

22 種類のナフトール AS 誘導体の構造と毒性との相関を計算化学の手法で検討した。化合物の毒性は、*Hydractinia echinata* のバイオアッセイによって決定した。Mlog(1/MRC_{50}) によって評価された毒性を目的変数として、量子力学計算などによって分子構造から求められるパラメータ（記述子）を説明変数として重回帰分析およびニューラルネットワークの手法を適用して相関モデルを構築した。変数選択は、遺伝的アルゴリズムを適用した。その結果、統計的に有意なモデルを構築でき、ナフトール AS 誘導体の毒性は、それらの分子の幾何学的な形状と化合物の極性に依存することが明らかになった。