

## Quantitative Structure-Activity Relationship Study of COX-2 Inhibitors

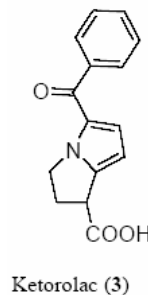
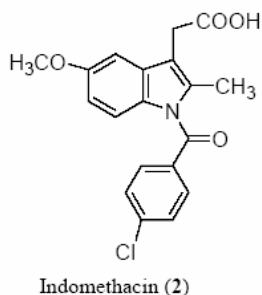
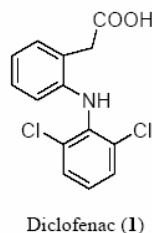
OLEG URSU, MIRCEA V. DIUDEA and SHIN-ICHI NAKAYMA

**ABSTRACT.** Quantitative Structure-Activity Relationship (*QSAR*) is a tool used with success to model biological activity of series of compounds. There is a serious limitation in the use of this method – it works well only on series of highly homogenous compounds. To furnish such sets of structures, several similarity classification schemes have been proposed: Naïve Bayesian classifier, artificial neural networks, molecular fingerprints, and maximum overlapping structures method, *MOS*. The most accurate classifier *MOS* is the most computationally demanding as well, and until recently its use was limited. Recent development of the algorithm introduced a series of heuristics and a new rapid clique detection procedure led to significant improvements in the speed of *MOS*, thus making it suitable for virtual screening of large databases of molecules. In the present study we make use of the *MOS* method to classify unknowns according to their structural features into classes of structurally homogeneous compounds suitable for further analysis by classical *QSAR* approaches.

### 1. INTRODUCTION

Cyclo-oxygenase enzyme is involved in inflammation and prostaglandin synthesis. The inhibition of the above synthesis by non-steroidal anti-inflammatory drugs (*NSAIDs*), makes these drugs ones of the most commonly used medications worldwide. These drugs are frequently used for the management of muscular-skeletal diseases and for (other caused) acute and chronic pain. Despite their clear efficacy, *NSAIDs* also cause adverse events, particularly gastrointestinal ulceration and altered renal function.

In the present study we focused on a method for finding new candidates with improved activity and hopefully reduced adverse effects. Three classes of *NSAIDs* were investigated: first class contains commonly used drugs, second and third classes contain new compounds still under investigations.<sup>2</sup> The study was carried out on a set of 82 structurally diverse *COX-2* inhibitors (see Figure 1).

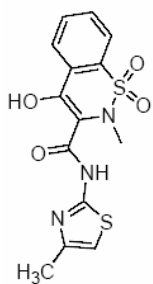


---

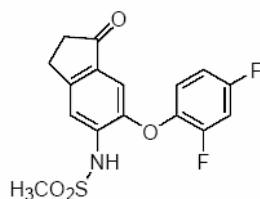
Received: 26.09.2004; In revised form: 17.01.2004

2000 *Mathematics Subject Classification.* 92E10

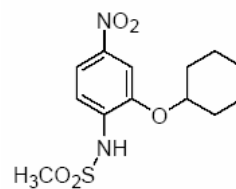
Key words and phrases. *QSAR, Maximum Overlapping Structures, Molecular Similarity*



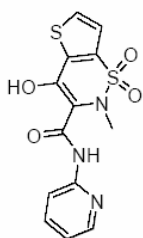
Meloxicam (4)



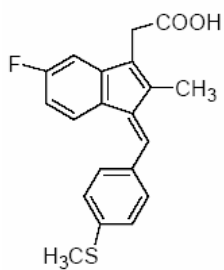
Flosulide (5)



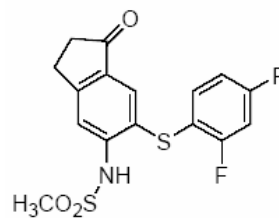
NS-398 (6)



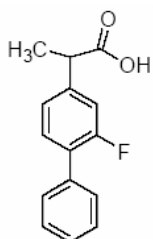
Tenoxicam (7)



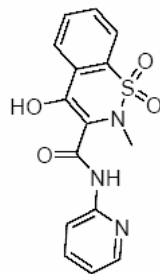
Sulindac sulphide (8)



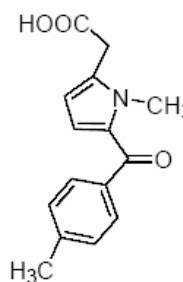
L-745337 (9)



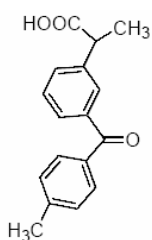
Flurbiprofen (10)



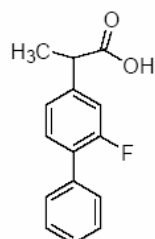
Piroxicam (11)



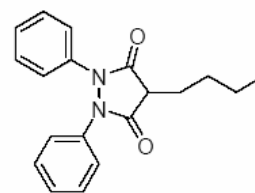
Tolmetin (12)



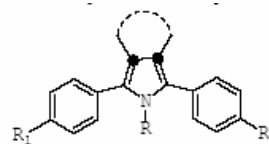
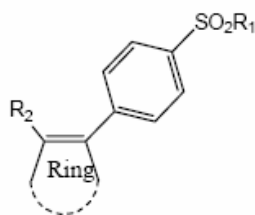
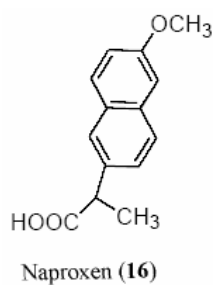
Ketoprofen (13)



Ibuprofen (14)



Phenyl butazone (15)



No.	Inhibitory activity $pIC_{50}$	No.	Inhibitory activity $pIC_{50}$	No.	Inhibitory activity $pIC_{50}$
1	7.30	29*	5.66	57*	7.78
2	6.34	30	4.76	58	7.67
3*	6.06	31	5.70	59	8.30
4	6.12	32	4.72	60*	7.38
5	6.12	33	5.33	61	7.00
6	6.33	34*	4.48	62	8.51
7	4.85	35	4.48	63	7.84
8*	4.98	36	4.48	64	6.00
9	5.01	37	4.48	65	9.15
10	5.19	38*	4.48	66*	8.54
11	5.05	39	4.48	67	7.00
12	5.15	40	4.48	68	7.00
13	5.97	41	4.48	69	8.58
14*	4.52	42	4.48	70*	7.10
15	4.52	43	4.48	71	8.35
16	4.13	44	4.48	72	6.15
17	7.22	45	4.00	73	5.00
18	6.00	46*	4.00	74*	8.79
19	6.30	47	0.00	75	6.92
20*	6.05	48	4.48	76	7.45
21	6.00	49	8.82	77	7.00
22	6.72	50*	8.48	78	7.00
23	7.52	51	6.00	79	7.00
24*	7.10	52	8.74	80	7.30
25	6.40	53	6.30	81	7.96
26	5.01	54	7.00	82	7.54
27	5.28	55	8.77		
28	4.87	56	6.30		

\*structures in external validation dataset

Figure 1. COX-2 Inhibitors molecular structure and inhibition activity  $pIC_{50}$  of the investigated dataset.

## 2. DATA SET ANALYSIS

All structures were sketched by using Hyperchem Molecular Modeling software package. Geometry optimization was performed by *MM+* force field prior to further optimization with semiempirical *PM3* Hamiltonian, available in Hyperchem. Molecular descriptors were generated by our *TOPOCLUJ* molecular topology software package<sup>3</sup> 3.0. A large set of molecular descriptors was generated (889). After separation of external validation set the remaining (training) set of molecules was classified by the *MOS* algorithm.<sup>5</sup> The *MOS* problem can be reduced at determining the maximum clique in modular product graph<sup>6-8</sup> (also known as correspondence graph, compatibility graph) -  $G_1 \diamond G_2$ . The concept was discovered on several occasions.<sup>1,7</sup> The modular product of two labeled line graphs  $G_1$  and  $G_2$  is defined on the vertex set  $V(G_1 \diamond G_2) = V(G_1) \times V(G_2)$  where the respective vertex labels are compatible, *i.e.*, both edges and the corresponding line graph vertices should have equivalent endpoint atoms. Two vertices  $(u_i, u_j)$  and  $(v_i, v_j)$ , where  $u_i, u_j$  represent vertices in  $L(G_1)$  while  $v_i, v_j$  denote vertices in  $L(G_2)$ , are connected in the modular product graph if:

$$\begin{aligned} &(u_i, u_j) \in E(G_1); (v_i, v_j) \in E(G_2) \text{ and } w(u_i, u_j) = w(v_i, v_j) \\ &\text{or } (u_i, u_j) \notin E(G_1) \text{ and } (v_i, v_j) \notin E(G_2) \end{aligned} \quad (1)$$

where  $w(u_i, u_j) = w(v_i, v_j)$  indicates that edge labels in line graphs are equivalent (edges in original graph are incident on the same atom type). This procedure is illustrated in Figure 2, on the molecular graphs  $G_1$  and  $G_2$  (hydrogen depleted) and their corresponding line graphs  $L(G_1)$  and  $L(G_2)$ .

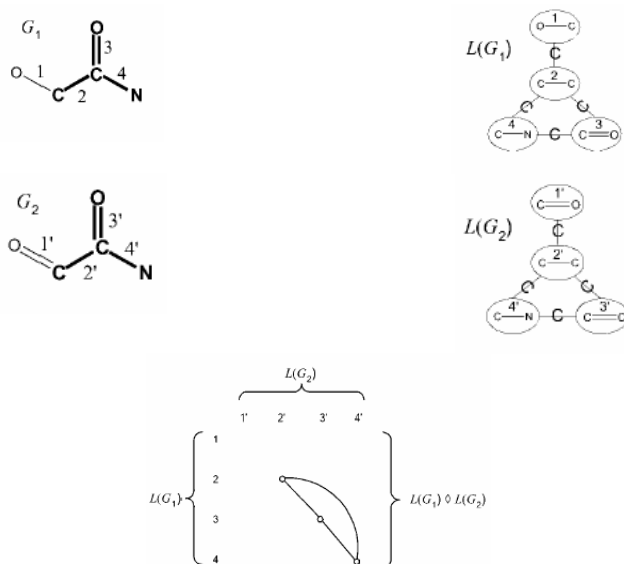
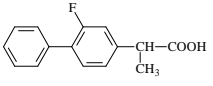
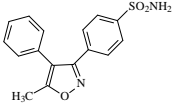
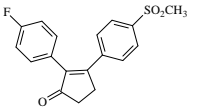
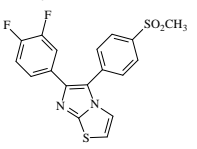
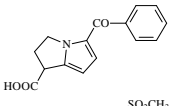
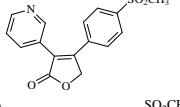
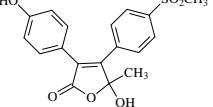
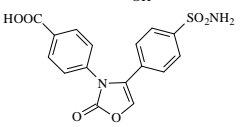
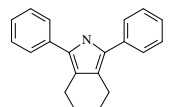
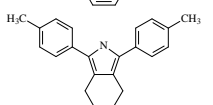
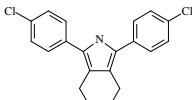
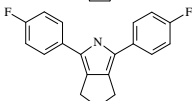
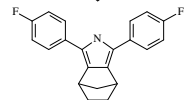
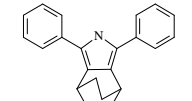
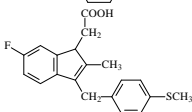


Figure 2. Graphs  $G_1$  and  $G_2$ , and their line graphs and modular product.

The molecules in the external validation dataset were classified using *MOS* algorithm. *MOS* was calculated for each molecule of the validation/prediction set with respect to all molecular structures from the (three) training subsets, their Tanimoto frequencies being listed in Table 1. All the unknowns were classified correctly, according to their structural features; this supports the high homology of each structural subset.

Table 1. Tanimoto frequencies for the external validation dataset

<i>No.</i>	<i>Structure</i>	<i>Subset 1</i> $\sum$ Tanimoto Frequencies	<i>Subset 2</i> $\sum$ Tanimoto Frequencies	<i>Subset 3</i> $\sum$ Tanimoto Frequencies
<b>cox14</b>		<b>0.05292</b>	<b>0.01653</b>	<b>0</b>
<b>cox20</b>		<b>0</b>	<b>0.17263</b>	<b>0.04948</b>
<b>cox24</b>		<b>0.02603</b>	<b>0.28659</b>	<b>0.01959</b>
<b>cox29</b>		<b>0</b>	<b>0.30414</b>	<b>0.18343</b>
<b>cox3</b>		<b>0.01205</b>	<b>0</b>	<b>0</b>
<b>cox34</b>		<b>0</b>	<b>0.23949</b>	<b>0.01163</b>
<b>cox38</b>		<b>0</b>	<b>0.20727</b>	<b>0.01034</b>
<b>cox46</b>		<b>0</b>	<b>0.18454</b>	<b>0.03335</b>
<b>cox50</b>		<b>0.01135</b>	<b>0.02781</b>	<b>0.70332</b>
<b>cox57</b>		<b>0.01073</b>	<b>0.02724</b>	<b>0.61362</b>

<b>cox60</b>		<b>0.01073</b>	<b>0.02634</b>	<b>0.62337</b>
<b>cox66</b>		<b>0.00975</b>	<b>0.02967</b>	<b>0.60989</b>
<b>cox70</b>		<b>0.01045</b>	<b>0.04205</b>	<b>0.70520</b>
<b>cox74</b>		<b>0.01073</b>	<b>0.08046</b>	<b>0.74977</b>
<b>cox8</b>		<b>0.01686</b>	<b>0</b>	<b>0</b>

Statistical analysis was carried out by *STATISTICA* software package.<sup>4</sup> Due to the large number of descriptors; we use Principal Component Analysis (*PCA*), a highly efficient method for reducing co-linearity of variables and dataset dimensionality. Factor coordinates projections for the three subsets are presented in Figure 3.

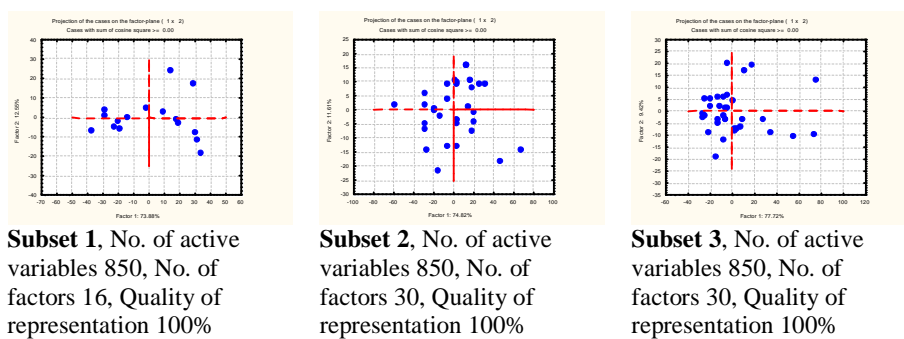
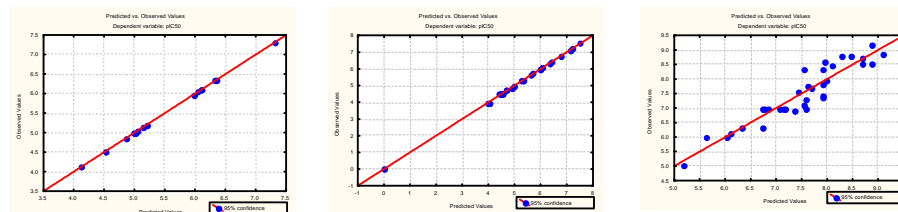


Figure 3. Case projection on factor coordinates

*PCA* analysis was carried out on the 850 descriptors and, due to structural diversity in each subset, a different number of *PCs* (Principal Components) is required to explain 100% variance. It is immediate, from the cases projection on two first factors coordinate's charts, that subset 3 has the highest homology. In model derivation, we used factor coordinates as independent variables in multiple linear regression analysis. Stepwise forward regression is the procedure of choice in our method because it automatically picks the independent variables which are the most statistically significant.

Excellent correlations are obtained for all the three subsets (Figure 4), thus proving the chosen independent variables have good correlation ability.



**Subset 1, Dependent:**  
pIC50  
Multiple R: 0.99997  
R<sup>2</sup>: 0.9999  
F = 1425.057  
No. of cases: 16  
Adjusted R<sup>2</sup>: 0.9992  
p = 0.020760  
Std. error of estimate:  
0.02356  
Intercept: 5.4775  
Std. Error: 0.00589  
No. of PC: 12

**Subset 2, Dependent:**  
pIC50  
Multiple R: 0.9896  
R<sup>2</sup>: 0.9794  
F = 26.15508  
No. of cases: 32  
Adjusted R<sup>2</sup>: 0.94195  
p = 0.000001  
Std. error of estimate:  
0.3259  
Intercept: 5.0750  
Std. Error: 0.05761  
No. of PC: 14

**Subset 3, Dependent:**  
pIC50  
Multiple R: 0.9389  
R<sup>2</sup>: 0.8816  
F = 8.94343  
No. of cases: 34  
Adjusted R<sup>2</sup>: 0.7831  
p = 0.000017  
Std. error of estimate:  
0.4661  
Intercept: 7.49176  
Std. Error: 0.07994  
No. of PC: 8

Figure 4. Statistical parameters for linear regressions *QSAR* models

External validation dataset was tested using the derived models for each cluster according to each molecule classification as shown above. The validation summary shows excellent prediction ability of the derived models thus proving that the present approach is suitable for *QSAR* investigations; validation plot and regression summary are presented in Figure 5.

Table 2. Predicted and residual values in the validation set.

Compound	Observed Value	Predicted Value	Residual	CV %
cox3	6.060	6.000	0.060	0.987
cox8	4.980	4.927	0.053	1.056
cox14	4.520	4.477	0.043	0.948
cox20	6.050	5.946	0.104	1.714
cox24	7.100	7.312	-0.212	2.991
cox29	5.660	5.475	0.185	3.274
cox34	4.480	4.638	-0.158	3.524
cox38	4.480	4.415	0.065	1.458
cox46	4.000	4.135	-0.135	3.381
cox50	8.480	8.026	0.454	5.357
cox57	7.780	7.553	0.227	2.914
cox60	7.380	7.852	-0.472	6.397
cox66	8.540	8.778	-0.238	2.790
cox70	7.100	7.477	-0.377	5.305
cox74	8.790	8.388	0.402	4.572

For each unknown in the external validation dataset the chosen regression model used for prediction is according to their *Tanimoto* frequencies in Table 1.

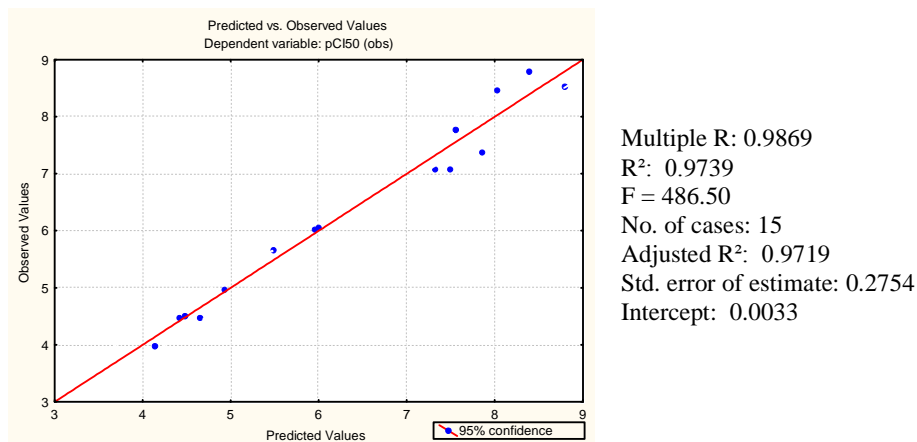


Figure 5. Predicted vs observed values, in the validation set.

### 3. CONCLUSIONS

A *QSAR* study on a dataset of 82 structurally diverse *COX-2* inhibitors was performed; we built up three simple and explicit *QSAR* models based on inhibitory activity of compounds in each training structural subset. The molecular structures in the external validation dataset were classified, using the *MOS* method, with high accuracy. The predicted values showed the excellent prediction ability of our derived models and correct classification of unknowns.

### REFERENCES

- [1] Raymond, J.W., Gardiner, E.J. and Willett P., *The Computer Journal*, **45** (2002), 631-645.
- [2] Chakraborti, A. K. and Thilagavathi R., *Internet Electronic Conference of Molecular Design*, (2003)
- [3] Diudea, M. V., Ursu O., *Indian J. Chem.*, **42 A** (2003), 1283-1294.
- [4] StatSoft, Inc. (2001). *STATISTICA* (data analysis software system), version 6. [www.statsoft.com](http://www.statsoft.com).
- [5] Raymond, J.W., Gardiner, E.J. and Willett P., *J. Chem. Inf. Comp. Sci.*, **42** (2002), 305-316.
- [6] Wood, D., *Oper. Res. Lett.*, **21** (1997), 211-217.
- [7] Bessonov, Y. E., *Vychisl. Sistemy*, **121** (1985), 3-22 (in Russian).
- [8] Chen, C. K., Yun, D. Y., *International Conference on Systems, Signals, Control, Computer*, Durban, South Africa, (1998).

BABEŞ - BOLYAI UNIVERSITY  
FACULTY OF CHEMISTRY AND CHEMICAL ENGINEERING  
ARANY JÁNOS 11, 40028, CLUJ, ROMANIA  
E-Mail Address: [diudea@chem.ubbcluj.ro](mailto:diudea@chem.ubbcluj.ro)

UNIVERSITY OF TSUKUBA  
RESEARCH CENTER FOR KNOWLEDGE COMMUNITIES  
1-2 KASUGA, TSUKUBA, IBARAKI, 305-8550, JAPAN