# Mutagenicity of Nitrated Polycyclic Aromatic Hydrocarbons: A OSAR Investigation 

Jyoti Singh ${ }^{1}$, Shalini Singh ${ }^{2}$, Basheerulla<br>Shaik ${ }^{1}$, Omar Deeb ${ }^{3}$, Neena Sohani ${ }^{4}$, Vijay K. Agrawal ${ }^{\mathbf{1}}$ and Padmakar V. Khadikar ${ }^{5, *}$<br>${ }^{1}$ OSAR and Computer Chemical Laboratories, A.P.S. University, Rewa 486 003, India<br>${ }^{2}$ Department of Chemistry, Bareilly College, Bareilly 243001, UP, India<br>${ }^{3}$ Faculty of Pharmacy, Al-Quds University, PO Box 20002, Jerusalem, Palestine<br>${ }^{4}$ Department of Chemistry, Pharmaceutical Chemistry and Industrial Chemistry, Christian Eminent Academy of Professional Education \& Research, LIG, Indore, India<br>${ }^{5}$ Research Division, Laxmi Fumigation and Pest Control, Pvt Ltd, 3, Khatipura, Indore 452 007, India<br>*Corresponding author: Padmakar V. Khadikar, pukhadikar@rediffmail.com


#### Abstract

Quantitative structure-activity relationship studies were performed to describe and predict the mutagenic activity of a set of 48 nitrated polycyclic aromatic hydrocarbons. From a larger pool of molecular descriptors (topological indices) we arrived at much a smaller set consisting of three correlating parameters. Such a variable selection is made using ncss software in that successive regressions were attempted using maximum- $\boldsymbol{R}^{2}$ method. The results are critically discussed using a variety of statistical parameters. Our results have shown that connectivity and shape type indices together with the distance-based Wiener index (W) play a dominating role in modelling of mutagenicity (logTA100). The predictive ability of the models is discussed on the basis of cross-validated parameters.


Key words: cheminformatics, ligand-based drug design, mutagenic activity, polycyclic hydrocarbons, quantitative structure-activity relationship, regression analysis

Received 26 April 2007, revised 22 December 2007 and accepted for publication 30 December 2007

Quantitative structure-activity relationship (OSAR) analysis can be defined as application of mathematical and statistical methods to the problem of finding QSAR models using experimental or calculated molecular descriptors of organic compounds acting as drugs (1-3). The goal of OSAR modelling was to establish a trend in
descriptor values which correlated with a trend in biological activity. All QSAR approaches implement directly or indirectly, a simple similarity principle, which for a long time has provided a foundation for the experimental medicinal chemistry: compounds with similar structures are expected to have similar biological activities.

Quantitative structure-activity relationship methods have been applied extensively in a wide range of scientific disciplines, including chemistry, biology and toxicology (4,5). In both drug discovery and environmental toxicology (6), OSAR models are now regarded as scientifically credible tools for predicting and classifying the biological activities of untested chemicals.

Polycyclic aromatic hydrocarbons (PAHs), in particular nitrated polycyclic aromatic hydrocarbons (Nitro-PAHs) are widespread environmental pollutants found in the exhaust fumes of gasoline and diesel combustion engines, in certain food products as a results of incomplete combustion and in general, in combustion source emissions (7-10). Nitro-PAHs have become of enormous concern because of their ubiquity in polluted air vapours, and because they are mostly associated with particular matter (PM).

Nitrated polycyclic aromatic hydrocarbons investigated from the mutation specificity end-point are quite limited as laboratory measurements are costly and time-consuming process; thus, prediction methods, such as QSAR modelling are needed to allow mutagenic activity estimation for a reliable risk assessment. It is worth to mention that the descriptors used earlier (11) were able to model mutagenic activity independent of the external prediction set composition. Consequently, the aim of the present investigation was to full model development on all of 48 Nitro-PAHs used in the present study. This will help us to compare our results with those of Gramatica et al. (11). We have relied upon OSAR methodology to derive statistically significant models that would relate the chemical structure of Nitro-PAHs to their mutagenic activity. A wide variety of descriptors (Table 3) have been used for OSAR analysis. These descriptors include distance- and connectivity-based topological indices together with shape indices. The list of the descriptors used in the present investigation is given in Appendix A.

## Material and Methods

## Mutagenicity

The structural details of 48 Nitro-PAHs used in the present study are given in Table 1 and their mutagenic activity ( $\operatorname{logTA100)}$ as adopted from the literature (11) are given in Table 2.

Table 1: Structural details of the nitrated polycyclic aromatic hydrocarbons used in the present study

| 1. 1,3,6,- Trinitropyrene | 2. 2,4,7-Trinitro-9-fluorenone | 3. 1,3-Dinitropyrene |
| :---: | :---: | :---: |
| 4. 1,6- Dinitropyrene | 5. 1,8- Dinitropyrene | 6. 2,7- Dinitro-9-fluorenone |
| 7. 1-Nitrofluoranthene | 8. 2-Nitroanthracene |  |
| 10. 3-Nitrofluoranthene | 11. 8-Nitrofluoranthene | 12. 1-Nitropyrene |
| 13. 7-Nitrofluoranthene | 14. 2-Nitronaphthalene | 15. 2-Nitrofluorene |
| 16. 2-Nitrophenanthrene | 17. 1-Nitronaphthalene | 18. 5-Nitroacenaphthalene |

## Singh et al.

Table 1: (Continued)

| 19. 1,3-Dinitrobenzene | 20. 1,3,5-Trinitrobenzene | 21. 4-Nitrotoulene |
| :---: | :---: | :---: |
| 22. 2,3-Dinitrotoulene | 23. 2,4-Dinitrotoulene | 24. 2,5-Dinitrotoulene |
| 25. 2,6-Dinitrotoulene | 26. 3,4-Dinitrotoulene | 27. 3,5-Dinitrotoulene |
| 28. 2, 3, 4-Trinitrotoulene | 29. 2,3,5-Trinitrotoulene | 30. 2,3,6-Trinitrotoulene |
| 31. 2,4,5-Trinitrotoulene | 32. 2,4,6-Trinitrotoulene | 33. 3,4,5-Trinitrotoulene |
| 34. 1-Me-2-nitronaphthalene | 35. 3-Me-2-nitronaphthalene | 36. 1,3-Dinitronaphthalene |

Table 1: (Continued)

| 37. 1,5-Dinitronaphthalene | 38. 1,8-Dinitronaphthalene | 39. 2,4,5,7-Tetra-nitro-9-fluorenone |
| :---: | :---: | :---: |
| 40. 2-Nitropyrene | 41. 1,3,6,8-Tetranitropyrene | 42. 5-Nitroquinolene |
| 43. 6-Nitroquinolene | 44. 2-Nitrocarbazole | 45. 3-Nitrocarbazole |
| 46. 4-Nitrocarbazole | 47. 9-Nitroanthracene | 48. 6-Nitrochrysene |

## Molecular descriptors

All the molecular descriptors viz. distance- and connectivity-based topological indices together with Jurs descriptors based on partial charges mapped on surface area (Tables 3-5) were computed using either the current version of dRAGON software ${ }^{\text {a }}$ or Karelson-CHEMAXON software ${ }^{\text {b }}$. For calculating these descriptors the structures were drawn using HYPERCHEM ${ }^{\text {c }}$ and ACD-LABS ${ }^{d}$ software. A total of 28 descriptors were chosen for the OSAR analysis (see Appendix A). The variable selection for multiple regression analysis has reduced this number to 10 , which then have been used for yielding statistically significant models.

## Chemometric methods

Multiple Linear Regression analysis and variable selection for multiple regression analysis were performed using ncss software ${ }^{e}$. The statistically significant models obtained are recorded in Table 6. It
is worthy to mention that the descriptors used earlier (11) were able to model mutagenic activity independent of the external prediction set composition. This allowed them to propose a full model development on 48 of the studied compounds. The models proposed in the present study are given in Table 6.

## Regression analysis

The regression analyses were performed using the maximum- $R^{2}$ method (12).

## Results and discussion

As stated earlier, in this paper we have proposed statistically significant QSAR models of mutagenic activity (logTA100) for a set of 48 Nitro-PAHs used in the present study. The present day statutory

## Singh et al.

Table 2: Name of the nitro-PAHs and their mutagenic activity (logTA100)

| Compound number | Chemicals | logTA100 |
| :---: | :---: | :---: |
| 1 | 1,3,6,-Trinitropyrene | 3.87 |
| 2 | 2,4,7-Trinitro-9-fluorenone | 2.27 |
| 3 | 1,3-Dinitropyrene | 4.63 |
| 4 | 1,6-Dinitropyrene | 4.09 |
| 5 | 1,8-Dinitropyrene | 4.74 |
| 6 | 2,7-Dinitro-9-fluorenone | 2.69 |
| 7 | 1-Nitrofluoranthene | 3.00 |
| 8 | 2-Nitroanthracene | 3.05 |
| 9 | 2,7-Dinitrofluorene | 1.27 |
| 10 | 3-Nitrofluoranthene | 3.31 |
| 11 | 8-Nitrofluoranthene | 2.60 |
| 12 | 1-Nitropyrene | 2.17 |
| 13 | 7-Nitrofluoranthene | 2.09 |
| 14 | 2-Nitronaphthalene | 0.37 |
| 15 | 2-Nitrofluorene | 1.08 |
| 16 | 2-Nitrophenanthrene | 1.79 |
| 17 | 1-Nitronaphthalene | 0.28 |
| 18 | 5-Nitroacenaphthalene | 0.97 |
| 19 | 1,3-Dinitrobenzene | -0.51 |
| 20 | 1,3,5-Trinitrobenzene | 0.72 |
| 21 | 4-Nitrotoulene | -2.10 |
| 22 | 2,3-Dinitrotoulene | -1.26 |
| 23 | 2,4-Dinitrotoulene | -1.29 |
| 24 | 2,5-Dinitrotoulene | -0.63 |
| 25 | 2,6-Dinitrotoulene | -1.34 |
| 26 | 3,4-Dinitrotoulene | -1.3 |
| 27 | 3,5-Dinitrotoulene | -0.72 |
| 28 | 2,3,4-Trinitrotoulene | 0.08 |
| 29 | 2,3,5-Trinitrotoulene | 0.46 |
| 30 | 2,3,6-Trinitrotoulene | 0.55 |
| 31 | 2,4,5-Trinitrotoulene | 1.12 |
| 32 | 2,4,6-Trinitrotoulene | 0.16 |
| 33 | 3,4,5-Trinitrotoulene | 1.01 |
| 34 | 1-Me-2-nitronaphthalene | 0.08 |
| 35 | 3-Me-2-nitronaphthalene | -0.70 |
| 36 | 1,3-Dinitronaphthalene | 0.86 |
| 37 | 1,5-Dinitronaphthalene | 0.91 |
| 38 | 1,8-Dinitronaphthalene | 1.12 |
| 39 | 2,4,5,7-Tetra-nitro-9-fluorenone | 2.46 |
| 40 | 2-Nitropyrene | 2.87 |
| 41 | 1,3,6,8-Tetranitropyrene | 3.18 |
| 42 | 5-Nitroquinolene | -0.70 |
| 43 | 6-Nitroquinolene | -1.05 |
| 44 | 2-Nitrocarbazole | -0.30 |
| 45 | 3-Nitrocarbazole | -1.00 |
| 46 | 4-Nitrocarbazole | -0.30 |
| 47 | 9-Nitroanthracene | 0.26 |
| 48 | 6-Nitrochrysene | 2.21 |

used in OSAR methodology is not just to reproduce known data, verified by fitting power $\left(R^{2}\right)$ but to predict the activity of chemicals not used in the development of OSAR model. This will help us to propose a full model, developed on all the 48 compounds used in the present study. This can be achieved by stepwise regression analysis using the method of maximum- $R^{2}(12-15)$. At this stage it is interesting to define the outlier and mention that the problem of identifying and dealing with outliers is a controversial issue and
one that seldom been addressed sufficiently in OSAR analysis. Statistically, if the residue, i.e. difference between the observed and calculated activity is more than two times standard deviation (discussion about such results is made at appropriate places in the section Discussion) then the compound is considered as an outlier. The reason for the removal of outliers and consequence of so doing (or indeed not doing) are poorly understood. As we know an outlier is a model that is not predicted well by a OSAR equation. Thus, we can use information that is generated about outliers to remove them iteratively from the OSAR equation, and then recalculate the equation until we are satisfied with the results. In all the proposed models there are no response outliers and no structurally influential chemicals, thus all belong to the chemical domain and the predicted data are reliable. As will be seen below, the molecular descriptors in our proposed models are very simple: they are 2D-topological descriptors and can be simply calculated from the molecular graph (hydrogen replaced molecular structure) without any conformational minimization or derived uncertainty on descriptors reproducibility.

Out of the several molecular descriptors used (Appendix A, Tables 3 and 4$)^{2} \chi^{v}$ alone gives a statistically significant model for modelling mutagenic activity ( $\log T A 100$ ). This model is found as below:

$$
\begin{equation*}
\log T A(100)=-3.7257+1.5520( \pm 0.1240)^{2} \chi^{v} \tag{1}
\end{equation*}
$$

$N=48, \mathrm{SE}=0.8153, R^{2}=0.7730, R^{2} A=0.7681, F=156.680$
Here and there after, $N$ is the number of compounds used, SE is the standard error of estimation, $R^{2}$ is the square of correlation coefficient $R, R^{2} A$ is adjusted $-R^{2}$ and $F$ is the Fisher's statistics. The positive coefficient of ${ }^{2} \chi^{v}$ indicates that the second-order branching and the presence of heteroatom are favourable for the exhibition of mutagenic activity $(\log T A 100)$.

Successive regression analysis yielded statistically significant (some superior and some inferior to the above one variable model) twovariable models (Table 6). Among these models, the one containing ${ }^{0} \chi^{v}$ and $W$ was found the best. This model is found as below:

$$
\begin{align*}
\operatorname{logTA}(100)= & -9.4626+1.6071( \pm 0.2203)^{0} \chi^{v} \\
& -0.0046( \pm 0.0012) W \tag{2}
\end{align*}
$$

$N=48, \mathrm{SE}=0.7563, R^{2}=0.8089, R^{2} A=0.8004, F=95.260$
The positive coefficient of ${ }^{0} \chi^{v}$ indicates that the number of atoms vis-à-vis molecular size and presence of heteroatom is favourable for the exhibition of $\log T A 100$. The Wiener index ( $M$ ) , accounts for the number of atoms, size, shape and branching. The negative coefficient of $W$ in the above equation appears to be due to its linear correlation with ${ }^{0} \chi^{v}$. This co-linearity aspect will be discussed elaborately in the following section. It is worthy to mention that the small coefficient on $W$ suggests that this descriptor contributes little to this model (this is applicable to all the following models in that $W$ is acting as one of the correlating parameters). However, from this model we positively infer that mutagenic activity is related to number of atoms, hetero-atoms, molecular size and thus

Table 3: Molecular descriptors of nitro-PAHs

| Compound number | $\operatorname{logTA100}$ | W | ${ }^{0} \chi$ | ${ }^{1} \chi$ | ${ }^{2} \chi$ | ${ }^{3} \chi$ | ${ }^{0} \chi^{v}$ | ${ }^{1} \chi^{v}$ | ${ }^{2} \chi^{v}$ | ${ }^{3} \chi^{v}$ | ${ }^{1} \chi^{\text {sh }}$ | ${ }^{2} \chi^{\text {sh }}$ | ${ }^{3} \chi^{\text {sh }}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 3.87 | 1262 | 17.8779 | 11.8967 | 11.5977 | 9.8099 | 11.7911 | 6.5162 | 4.9421 | 3.6758 | 15.1466 | 5.0464 | 2.0063 |
| 2 | 2.27 | 1078 | 16.8863 | 10.8074 | 10.6285 | 8.7212 | 10.6993 | 5.7203 | 4.2712 | 3.0181 | 14.7158 | 4.9132 | 2.1367 |
| 3 | 4.63 | 878 | 15.4303 | 10.5753 | 10.1396 | 8.7701 | 10.5274 | 5.9275 | 4.4614 | 3.3255 | 12.7560 | 4.2662 | 1.6555 |
| 4 | 4.09 | 905 | 15.4303 | 10.5753 | 10.1277 | 8.8175 | 10.5274 | 5.9275 | 4.4614 | 3.3255 | 12.7560 | 4.2662 | 1.6555 |
| 5 | 4.74 | 905 | 15.4303 | 10.5753 | 10.1277 | 8.8175 | 10.5274 | 5.9275 | 4.4614 | 3.3255 | 12.7560 | 4.2662 | 1.6555 |
| 6 | 2.69 | 770 | 14.4387 | 9.4860 | 9.1488 | 7.7775 | 9.4356 | 5.1316 | 3.7905 | 2.6678 | 12.2733 | 4.1194 | 1.7661 |
| 7 | 3.00 | 604 | 12.9827 | 9.2540 | 8.6837 | 7.7434 | 9.2637 | 5.3387 | 3.9807 | 2.9818 | 10.4068 | 3.4936 | 1.3020 |
| 8 | 3.05 | 500 | 11.8280 | 8.2371 | 7.6133 | 6.4007 | 8.2637 | 4.5887 | 3.2307 | 2.1693 | 9.9302 | 3.6435 | 1.6738 |
| 9 | 1.27 | 693 | 13.5685 | 9.0585 | 8.7178 | 7.2046 | 9.0274 | 4.9275 | 3.5864 | 2.4637 | 11.6392 | 4.0440 | 1.8368 |
| 10 | 3.31 | 601 | 12.9827 | 9.2708 | 8.5972 | 7.7724 | 9.2637 | 5.3387 | 3.9807 | 3.0377 | 10.4068 | 3.4936 | 1.2537 |
| 11 | 2.60 | 619 | 12.9827 | 9.2540 | 8.6718 | 7.7961 | 9.2637 | 5.3387 | 3.9807 | 2.9818 | 10.4068 | 3.4936 | 1.3020 |
| 12 | 2.17 | 596 | 12.9827 | 9.2540 | 8.6696 | 7.7777 | 9.2637 | 5.3387 | 3.9807 | 2.9752 | 10.4068 | 3.4936 | 1.3020 |
| 13 | 2.09 | 619 | 12.9827 | 9.2540 | 8.6718 | 7.7961 | 9.2637 | 5.3387 | 3.9807 | 2.9818 | 10.4068 | 3.4936 | 1.3020 |
| 14 | 0.37 | 238 | 9.2591 | 6.2708 | 5.6217 | 4.5230 | 6.2637 | 3.3387 | 2.2307 | 1.4193 | 7.8859 | 2.9531 | 1.4293 |
| 15 | 1.08 | 414 | 11.1209 | 7.7540 | 7.1851 | 6.1479 | 7.7637 | 4.3387 | 3.1057 | 2.1693 | 9.2421 | 3.2500 | 1.3702 |
| 16 | 1.79 | 492 | 11.8280 | 8.2540 | 7.5268 | 6.4506 | 8.2637 | 4.5887 | 3.2307 | 2.2318 | 9.9302 | 3.6435 | 1.5691 |
| 17 | 0.28 | 226 | 9.2591 | 6.2876 | 5.5471 | 4.4864 | 6.2637 | 3.3387 | 2.2307 | 1.4752 | 7.8859 | 2.9531 | 1.2938 |
| 18 | 0.97 | 317 | 10.4138 | 7.2708 | 6.7548 | 5.8807 | 7.2637 | 4.0887 | 2.9807 | 2.1627 | 8.3200 | 2.7448 | 1.0467 |
| 19 | -0.51 | 197 | 9.1378 | 5.6090 | 5.1747 | 3.6409 | 5.5274 | 2.6775 | 1.7114 | 0.9637 | 8.4367 | 3.1592 | 1.9121 |
| 20 | 0.72 | 354 | 11.5854 | 6.9135 | 6.7193 | 4.6449 | 6.7911 | 3.2662 | 2.1921 | 1.2581 | 10.9775 | 3.9580 | 2.4524 |
| 21 | -2.10 | 120 | 7.5604 | 4.6983 | 4.2639 | 3.0033 | 4.7637 | 2.3387 | 1.4807 | 0.7943 | 6.8938 | 2.5772 | 1.6267 |
| 22 | -1.26 | 228 | 10.0080 | 6.0365 | 5.6374 | 3.9744 | 6.0274 | 2.9275 | 1.9614 | 1.1946 | 9.4281 | 3.3783 | 1.7585 |
| 23 | -1.29 | 240 | 10.0080 | 6.0197 | 5.7023 | 4.0977 | 6.0274 | 2.9275 | 1.9614 | 1.1446 | 9.4281 | 3.3783 | 1.9228 |
| 24 | -0.63 | 246 | 10.0080 | 6.0197 | 5.7023 | 4.0753 | 6.0274 | 2.9275 | 1.9614 | 1.1446 | 9.4281 | 3.3783 | 1.9228 |
| 25 | -1.34 | 234 | 10.0080 | 6.036581 | 5.6057 | 4.1382 | 6.0274 | 2.9275 | 1.9614 | 1.2005 | 9.4281 | 3.3783 | 1.7585 |
| 26 | -1.3 | 234 | 10.0080 | 6.019744 | 5.7339 | 3.9298 | 6.0274 | 2.9275 | 1.9614 | 1.1387 | 9.4281 | 3.3783 | 1.9228 |
| 27 | -0.72 | 240 | 10.0080 | 6.0029 | 5.8204 | 3.9013 | 6.0274 | 2.9275 | 1.9614 | 1.0887 | 9.4281 | 3.3783 | 2.1630 |
| 28 | 0.08 | 393 | 12.4556 | 7.3579 | 7.1052 | 4.9471 | 7.2911 | 3.5162 | 2.4421 | 1.5390 | 11.9719 | 4.1820 | 2.1180 |
| 29 | 0.46 | 408 | 12.4556 | 7.3411 | 7.1820 | 4.9698 | 7.2911 | 3.5162 | 2.4421 | 1.4890 | 11.9719 | 4.1820 | 2.3010 |
| 30 | 0.55 | 405 | 12.4556 | 7.3579 | 7.0735 | 5.0925 | 7.2911 | 3.5162 | 2.4421 | 1.5449 | 11.9719 | 4.1820 | 2.1180 |
| 31 | 1.12 | 411 | 12.4556 | 7.3411 | 7.1723 | 5.0380 | 7.2911 | 3.5162 | 2.4421 | 1.4890 | 11.9719 | 4.1820 | 2.3010 |
| 32 | 0.16 | 408 | 12.4556 | 7.3411 | 7.1503 | 5.1559 | 7.2911 | 3.5162 | 2.4421 | 1.4949 | 11.9719 | 4.1820 | 2.3010 |
| 33 | 1.01 | 396 | 12.4556 | 7.3411 | 7.2136 | 4.8202 | 7.2911 | 3.5162 | 2.4421 | 1.4831 | 11.9719 | 4.1820 | 2.3010 |
| 34 | 0.08 | 280 | 10.1293 | 6.6983 | 6.0527 | 5.0456 | 6.7637 | 3.5887 | 2.4807 | 1.6627 | 8.8405 | 3.1953 | 1.4135 |
| 35 | -0.70 | 284 | 10.1293 | 6.6815 | 6.1611 | 4.9031 | 6.7637 | 3.5887 | 2.4807 | 1.6002 | 8.8405 | 3.1953 | 1.5131 |
| 36 | 0.86 | 403 | 11.7067 | 7.5922 | 7.0917 | 5.4973 | 7.5274 | 3.9275 | 2.7114 | 1.7696 | 10.3415 | 3.7619 | 1.8069 |
| 37 | 0.91 | 400 | 11.7067 | 7.6090 | 7.0052 | 5.5122 | 7.5274 | 3.9275 | 2.7114 | 1.8255 | 10.3415 | 3.7619 | 1.6999 |
| 38 | 1.12 | 391 | 11.7067 | 7.6090 | 7.0149 | 5.4704 | 7.5274 | 3.9275 | 2.7114 | 1.8255 | 10.3415 | 3.7619 | 1.6999 |
| 39 | 2.46 | 1443 | 19.3339 | 12.1288 | 12.1082 | 9.6705 | 11.9630 | 6.3091 | 4.7519 | 3.3684 | 17.1820 | 5.7100 | 2.5236 |
| 40 | 2.87 | 614 | 12.9827 | 9.2371 | 8.7561 | 7.7418 | 9.2637 | 5.3387 | 3.9807 | 2.9193 | 10.4068 | 3.4936 | 1.3664 |
| 41 | 3.18 | 1694 | 20.3255 | 13.2181 | 13.0677 | 10.8023 | 13.0548 | 7.1050 | 5.4228 | 4.0261 | 17.5676 | 5.8320 | 2.3739 |
| 42 | -0.70 | 226 | 9.2591 | 6.2876 | 5.5471 | 4.4864 | 6.2109 | 3.2859 | 2.1779 | 1.4290 | 7.8193 | 2.9124 | 1.2709 |
| 43 | -1.05 | 238 | 9.2591 | 6.2708 | 5.6217 | 4.5230 | 6.2109 | 3.2859 | 2.1779 | 1.3731 | 7.8193 | 2.9124 | 1.4047 |
| 44 | -0.30 | 414 | 11.1209 | 7.7540 | 7.1851 | 6.1479 | 7.7109 | 4.2859 | 3.0397 | 2.1033 | 9.2051 | 3.2291 | 1.3593 |
| 45 | -1.00 | 408 | 11.1209 | 7.7540 | 7.1851 | 6.1547 | 7.7109 | 4.2859 | 3.0397 | 2.1033 | 9.2051 | 3.2291 | 1.3593 |
| 46 | -0.30 | 387 | 11.1209 | 7.7708 | 7.1202 | 6.0751 | 7.7109 | 4.2859 | 3.0397 | 2.1592 | 9.2051 | 3.2291 | 1.2737 |
| 47 | 0.26 | 452 | 11.8280 | 8.2708 | 7.4738 | 6.2998 | 8.2637 | 4.5887 | 3.2307 | 2.2811 | 9.9302 | 3.6435 | 1.4917 |
| 48 | 2.21 | 810 | 14.3969 | 10.2540 | 9.3691 | 8.3139 | 10.2637 | 5.8387 | 4.2307 | 3.1002 | 11.9984 | 4.3639 | 1.7155 |

mutagenic activity increases with number of rings and the number of nitro-groups in the set of 48 Nitro-PAHs used. Also, the shape of the molecules is directly related to the mutagenic activity (logTA100). Consequently, we can argue that Nitro-PAHs with a less linear, more round (circular) shape would be most active. Finally, this two-variable model alone accounts for $80 \%$ variation in the mutagenic activity (logTA100) of the Nitro-PAHs used.
rior to the two-variable model discussed above. Out of the superior models, the model containing ${ }^{3} \chi^{v},{ }^{3} \chi^{\text {shape }}$ and $W$ as the correlating parameters is found to be the best:

$$
\begin{align*}
\operatorname{logTA}(100)= & -8.3701+3.8140( \pm 0.5365)^{3} \chi^{v} \\
& +2.3907( \pm 0.5796)^{3} \chi^{\text {shape }}-0.0053( \pm 0.0014) W
\end{align*}
$$

$N=48, \mathrm{SE}=0.7142, R^{2}=0.8334, R^{2} A=0.8220, F=73.365$

Further stepwise regression analysis gave 14 three-variable models (Table 6). Here also, some models are superior and some are infe-

## Singh et al.

Table 4: Molecular descriptors of nitro-PAHs

| Compound number | TMSA | PPSA1 | PPSA2 | PPSA3 | PNSA1 | PNSA2 | PNSA-3 | DPSA1 | DPSA2 | DPSA3 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 528.0454 | 221.2820 | 81.8331 | 2.8269 | 306.7634 | -113.4453 | -18.9088 | -85.4814 | 195.2784 | 21.7357 |
| 2 | 548.2105 | 182.6731 | 88.1688 | 4.0373 | 365.5374 | -176.4300 | -24.3532 | -182.8642 | 264.5988 | 28.3906 |
| 3 | 496.5130 | 291.0332 | 71.4655 | 2.0126 | 205.4797 | -50.4572 | -12.6143 | 85.5534 | 121.9227 | 14.6269 |
| 4 | 497.0392 | 288.3523 | 71.8113 | 2.8512 | 208.6868 | -51.9714 | -12.9928 | 79.6655 | 123.7828 | 15.8440 |
| 5 | 497.0392 | 288.3523 | 71.8113 | 2.8512 | 208.6868 | -51.9714 | -12.9928 | 79.6655 | 123.7828 | 15.8440 |
| 6 | 482.1913 | 215.1559 | 78.2443 | 3.6943 | 267.0354 | -97.1110 | -18.5608 | -51.8794 | 175.3554 | 22.2551 |
| 7 | 440.3387 | 327.2741 | 40.6836 | 1.1017 | 113.0646 | -14.0551 | -7.0275 | 214.2095 | 54.7388 | 8.1292 |
| 8 | 414.7085 | 300.4658 | 37.3296 | 1.5113 | 114.2427 | -14.1934 | -7.0967 | 186.2231 | 51.5231 | 8.6080 |
| 9 | 477.5901 | 246.6503 | 61.2585 | 2.8005 | 230.9398 | -57.3566 | -14.3391 | 15.7105 | 118.6152 | 17.1396 |
| 10 | 433.1693 | 328.3514 | 40.9246 | 1.3572 | 104.8179 | -13.0641 | -6.5320 | 223.5335 | 53.9888 | 7.8893 |
| 11 | 436.5806 | 323.5159 | 40.1956 | 1.3630 | 113.0646 | -14.0478 | -7.0239 | 210.4513 | 54.2435 | 8.3870 |
| 12 | 405.9971 | 305.5643 | 38.0848 | 1.2068 | 100.4327 | -12.5177 | -6.2588 | 205.1316 | 50.6025 | 7.4656 |
| 13 | 470.4548 | 353.0050 | 43.8595 | 1.7019 | 117.4497 | -14.5926 | -7.2963 | 235.5553 | 58.4522 | 8.9983 |
| 14 | 364.5621 | 247.1124 | 30.7005 | 1.6392 | 117.4497 | -14.5916 | -7.2958 | 129.6627 | 45.2922 | 8.9350 |
| 15 | 420.3464 | 305.6782 | 37.9773 | 1.5683 | 114.6681 | -14.2463 | -7.1231 | 191.0101 | 52.2236 | 8.6915 |
| 16 | 427.1361 | 310.1118 | 38.5285 | 1.5517 | 117.024 | -14.5391 | -7.2695 | 193.0875 | 53.0677 | 8.8213 |
| 17 | 340.8432 | 237.2033 | 29.5576 | 1.5087 | 103.6398 | -12.9144 | -6.4572 | 133.5635 | 42.4720 | 7.9659 |
| 18 | 365.2337 | 259.9903 | 32.4033 | 1.3038 | 105.2433 | -13.1167 | -6.5583 | 154.7469 | 45.5201 | 7.8622 |
| 19 | 353.5064 | 125.3482 | 30.5983 | 2.5560 | 228.1581 | -55.6949 | -13.923 | -102.8099 | 86.2933 | 16.4797 |
| 20 | 419.2036 | 90.7108 | 32.6889 | 2.5671 | 328.4927 | -118.3770 | -19.7295 | -237.7819 | 151.066 | 22.2966 |
| 21 | 343.8944 | 228.0482 | 28.3152 | 2.0089 | 115.8462 | -14.3838 | -7.1919 | 112.2019 | 42.6991 | 9.2009 |
| 22 | 353.6006 | 188.6997 | 45.2197 | 2.8089 | 164.9008 | -39.5166 | -9.8791 | 23.7988 | 84.7364 | 12.6880 |
| 23 | 383.2782 | 172.0388 | 42.0914 | 2.6214 | 211.2393 | -51.6823 | -12.9192 | -39.2005 | 93.7738 | 15.5406 |
| 24 | 392.5380 | 174.9827 | 43.0742 | 3.1412 | 217.5552 | -53.5540 | -13.3872 | -42.5725 | 96.6283 | 16.5285 |
| 25 | 360.8420 | 166.4887 | 40.8661 | 2.9799 | 194.3533 | -47.7058 | -11.9264 | -27.8646 | 88.5719 | 14.9064 |
| 26 | 371.4213 | 197.1938 | 47.1629 | 2.6612 | 174.2274 | -41.6700 | -10.4176 | 22.9663 | 88.8329 | 13.0788 |
| 27 | 392.3679 | 164.2097 | 40.1250 | 2.1529 | 228.1581 | -55.7510 | -13.9377 | -63.9484 | 95.8761 | 16.0907 |
| 28 | 400.0700 | 174.4317 | 61.3872 | 3.3732 | 225.6383 | -79.4082 | -13.2457 | -51.2066 | 140.7955 | 16.6190 |
| 29 | 421.2404 | 148.4127 | 53.1105 | 2.9865 | 272.8277 | -97.6333 | -16.3034 | -124.4149 | 150.7439 | 19.2899 |
| 30 | 417.2820 | 153.3228 | 54.9840 | 3.7410 | 263.9592 | -94.6600 | -15.7925 | -110.6364 | 149.6440 | 19.5335 |
| 31 | 415.3159 | 148.1496 | 52.9989 | 2.9446 | 267.1662 | -95.5757 | -15.9362 | -119.0166 | 148.5746 | 18.8809 |
| 32 | 427.1938 | 117.4196 | 42.5067 | 2.8481 | 309.7741 | -112.1404 | -18.6872 | -192.3544 | 154.6472 | 21.5354 |
| 33 | 411.4517 | 168.4692 | 59.2116 | 2.7053 | 242.9825 | -85.4008 | -14.2563 | -74.5133 | 144.6125 | 16.9617 |
| 34 | 383.1093 | 280.9748 | 35.0017 | 1.6855 | 102.1344 | -12.7231 | -6.3615 | 178.8403 | 47.7249 | 8.0471 |
| 35 | 380.6235 | 284.4450 | 35.4384 | 1.7246 | 96.1785 | -11.9827 | -5.9913 | 188.2665 | 47.4211 | 7.7160 |
| 36 | 421.0978 | 207.9276 | 50.9371 | 2.2077 | 213.1701 | -52.2213 | -13.0543 | -5.2425 | 103.1585 | 15.2621 |
| 37 | 387.1201 | 178.9896 | 44.3906 | 2.4746 | 208.1305 | -51.6177 | -12.9044 | -29.1409 | 96.0083 | 15.3790 |
| 38 | 372.6779 | 226.4303 | 55.7606 | 2.8054 | 146.2476 | -36.0148 | -9.0037 | 80.1826 | 91.7755 | 11.8091 |
| 39 | 575.2923 | 149.6392 | 89.9100 | 3.9509 | 425.6531 | -255.7515 | -27.835 | -276.0137 | 345.6615 | 31.7861 |
| 40 | 459.4683 | 343.6221 | 42.7143 | 1.4010 | 115.8462 | -14.4003 | -7.2001 | 227.7759 | 57.1147 | 8.6012 |
| 41 | 601.6254 | 190.6658 | 93.4982 | 3.6496 | 410.9595 | -201.5252 | -25.1906 | -220.2937 | 295.0235 | 28.8403 |
| 42 | 340.5676 | 213.6021 | 44.2264 | 3.6298 | 126.9655 | -26.2882 | -8.3269 | 86.6366 | 70.5147 | 11.9567 |
| 43 | 356.1587 | 220.3041 | 45.7521 | 3.7909 | 135.8545 | -28.2139 | -8.9678 | 84.4496 | 73.9660 | 12.7587 |
| 44 | 409.5616 | 242.8281 | 50.6530 | 2.2954 | 166.7334 | -34.7799 | -11.4930 | 76.0946 | 85.4329 | 13.7885 |
| 45 | 418.3934 | 248.4528 | 52.0946 | 2.6462 | 169.9405 | -35.6324 | -11.7873 | 78.5123 | 87.7271 | 14.4336 |
| 46 | 396.8905 | 251.1336 | 52.4814 | 2.4393 | 145.7568 | -30.4599 | -10.0674 | 105.3768 | 82.9414 | 12.5067 |
| 47 | 389.6124 | 294.8410 | 36.8691 | 1.1365 | 94.7713 | -11.8509 | -5.9254 | 200.0697 | 48.7200 | 7.0620 |
| 48 | 448.4309 | 349.7980 | 43.6189 | 1.0330 | 98.6329 | -12.2992 | -6.1496 | 251.1651 | 55.9182 | 7.1827 |

The positive coefficients of both ${ }^{3} \chi^{\vee}$ and ${ }^{3} \chi^{\text {shape }}$ favours the contribution of shape, size, etc. in exhibiting the mutagenic activity (logTA100). A perusal of Table 6 shows that for four- and five-variable models only a slight improvement in the statistics occurs. This means that the three-variable model discussed above can be considered as the most appropriate model for estimating mutagenic activity $(\log T A 100)$ of the 48 Nitro-PAHs used in the present study.

However, we performed still higher parametric regressions and go up to 10 -parametric model. Consistent improvement in statistics
occurred during higher parametric regression analyses. We thought this improvement is probably due to increase in the correlating parameters. However, we also observed increase in $R^{2} A$ when we proceed up to 10 -parametic model. This clearly means that added parameters in succession contribute favourably and significantly to the arrived models. Furthermore, the rule of thumb (16) also suggests that the 10 -parametric model is allowed.

Before proceeding for further investigation it is important to make our self familiar with the rule of thumb (16) that will help us

Table 5: Molecular descriptors of nitro-PAHs

| Compound number | FPSA1 | FPSA2 | FPSA3 | FNSA1 | FNSA2 | FNSA3 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.4190 | 0.1549 | 0.0053 | 0.5809 | -0.2148 | -0.0358 |
| 2 | 0.3332 | 0.1608 | 0.0073 | 0.6667 | -0.3218 | -0.0444 |
| 3 | 0.5861 | 0.1439 | 0.0040 | 0.4138 | -0.1016 | -0.0254 |
| 4 | 0.5801 | 0.1444 | 0.0057 | 0.4198 | -0.1045 | -0.0261 |
| 5 | 0.5801 | 0.1444 | 0.0057 | 0.4198 | -0.1045 | -0.0261 |
| 6 | 0.4462 | 0.1622 | 0.0076 | 0.5537 | -0.2013 | -0.0384 |
| 7 | 0.7432 | 0.0923 | 0.0025 | 0.2567 | -0.0319 | -0.0159 |
| 8 | 0.7245 | 0.0900 | 0.0036 | 0.2754 | -0.0342 | -0.0171 |
| 9 | 0.5164 | 0.1282 | 0.0058 | 0.4835 | -0.1200 | -0.0300 |
| 10 | 0.7580 | 0.0944 | 0.0031 | 0.2419 | -0.0301 | -0.0150 |
| 11 | 0.7410 | 0.0920 | 0.0031 | 0.2589 | -0.0321 | -0.0160 |
| 12 | 0.7526 | 0.0938 | 0.0029 | 0.2473 | -0.0308 | -0.0154 |
| 13 | 0.7503 | 0.0932 | 0.0036 | 0.2496 | -0.0310 | -0.0155 |
| 14 | 0.6778 | 0.0842 | 0.0044 | 0.3221 | -0.0400 | -0.0200 |
| 15 | 0.7272 | 0.0903 | 0.0037 | 0.2727 | -0.0338 | -0.0169 |
| 16 | 0.7260 | 0.0902 | 0.0036 | 0.2739 | -0.0340 | -0.0170 |
| 17 | 0.6959 | 0.0867 | 0.0044 | 0.3040 | -0.0378 | -0.0189 |
| 18 | 0.7118 | 0.0887 | 0.0035 | 0.2881 | -0.0359 | -0.0179 |
| 19 | 0.3545 | 0.0865 | 0.0072 | 0.6454 | -0.1575 | -0.0393 |
| 20 | 0.2163 | 0.0779 | 0.0061 | 0.7836 | -0.2823 | -0.0470 |
| 21 | 0.6631 | 0.0823 | 0.0058 | 0.3368 | -0.0418 | -0.0209 |
| 22 | 0.5336 | 0.1278 | 0.0079 | 0.4663 | -0.1117 | -0.0279 |
| 23 | 0.4488 | 0.1098 | 0.0068 | 0.5511 | -0.1348 | -0.0337 |
| 24 | 0.4457 | 0.1097 | 0.0080 | 0.5542 | -0.1364 | -0.0341 |
| 25 | 0.4613 | 0.1132 | 0.0082 | 0.5386 | -0.1322 | -0.0330 |
| 26 | 0.5309 | 0.1269 | 0.0071 | 0.4690 | -0.1121 | -0.0280 |
| 27 | 0.4185 | 0.1022 | 0.0054 | 0.5814 | -0.1420 | -0.0355 |
| 28 | 0.4360 | 0.1534 | 0.0084 | 0.5639 | -0.1984 | -0.0331 |
| 29 | 0.3523 | 0.1260 | 0.0070 | 0.6476 | -0.2317 | -0.0387 |
| 30 | 0.3674 | 0.1317 | 0.0089 | 0.6325 | -0.2268 | -0.0378 |
| 31 | 0.3567 | 0.1276 | 0.0070 | 0.6432 | -0.2301 | -0.0383 |
| 32 | 0.2748 | 0.0995 | 0.0066 | 0.7251 | -0.2625 | -0.0437 |
| 33 | 0.4094 | 0.1439 | 0.0065 | 0.5905 | -0.2075 | -0.0346 |
| 34 | 0.7334 | 0.0913 | 0.0044 | 0.2665 | -0.0332 | -0.0166 |
| 35 | 0.7473 | 0.0931 | 0.0045 | 0.2526 | -0.0314 | -0.0157 |
| 36 | 0.4937 | 0.1209 | 0.0052 | 0.5062 | -0.1240 | -0.0310 |
| 37 | 0.4623 | 0.1146 | 0.0063 | 0.5376 | -0.1333 | -0.0333 |
| 38 | 0.6075 | 0.1496 | 0.0075 | 0.3924 | -0.0966 | -0.0241 |
| 39 | 0.2601 | 0.1562 | 0.0068 | 0.7398 | -0.4445 | -0.0483 |
| 40 | 0.7478 | 0.0929 | 0.0030 | 0.2521 | -0.0313 | -0.0156 |
| 41 | 0.3169 | 0.1554 | 0.0060 | 0.6830 | -0.3349 | -0.0418 |
| 42 | 0.6271 | 0.1298 | 0.0106 | 0.3728 | -0.0771 | -0.0244 |
| 43 | 0.6185 | 0.1284 | 0.0106 | 0.3814 | -0.0792 | -0.0251 |
| 44 | 0.5928 | 0.1236 | 0.0056 | 0.4071 | -0.0849 | -0.0280 |
| 45 | 0.5938 | 0.1245 | 0.0063 | 0.4061 | -0.0851 | -0.0281 |
| 46 | 0.6327 | 0.1322 | 0.0061 | 0.3672 | -0.0767 | -0.0253 |
| 47 | 0.7567 | 0.0946 | 0.0029 | 0.2432 | -0.0304 | -0.0152 |
| 48 | 0.7800 | 0.0972 | 0.0023 | 0.2199 | -0.0274 | -0.0137 |

whether or not we can undergo higher parametric regression analysis. The technique that has been most used in OSAR is linear multiple regression, which employs the least-squares method to find the equation of 'best fit' of biological activity with a given combination of parameters (descriptors). The limitations and some common pitfalls of multiple regression analysis were pointed out by Tute (16). Accordingly, there must be a sufficient number of compounds included in the analysis to enable statistical significance to be reached, despite evitable errors in measurement. A
rule of thumb was evolved (16) that at least five data points (compounds) should be included for every parameter in the equation. The parameter themselves should be well 'spread'. Thus, looking to the number of compounds used, 48, and in accordance with the rule of thumb we can at the most go for 10 -parametric regression analysis. Such nine- and 10 -parametric models are mentioned below:
(i) Nine-variable model
$\log T A(100)=-17.1337-7.3005( \pm 1.4057)^{1} \chi^{\text {shape }}$
$+4.8927( \pm 1.0820)^{3} \chi^{\text {shape }}+0.0390( \pm 0.0143)$ PNSA- 1
$-0.0087( \pm 0.0016) W+0.8960( \pm 0.2170)$ PNSA-2
$+10.8457( \pm 2.1396)^{0} \chi-4.4003( \pm 1.3593)^{1} \chi$
$+4.3001( \pm 1.7104)^{3} \chi-9.1531( \pm 3.2112)^{2} \chi^{v}$
$N=48, S E=0.5540, R^{2}=0.9134, R^{2} A=0.8929, F=44.549$
(ii) 10 -variable model

$$
\begin{align*}
\log T \mathrm{~A}(100)= & -18.7580-3.3543( \pm 2.2555)^{3} \chi^{v} \\
& -7.7489( \pm 1.4471)^{1} \chi^{\text {shape }}+4.1156( \pm 1.3419)^{3} \chi^{\text {shape }} \\
& +0.0477( \pm 0.0164) \text { PNSA-1 }-0.0093( \pm 0.0011) W \\
& +0.9277( \pm 0.2539) \text { PNSA- } 2+12.6777( \pm 2.4256)^{0} \chi \\
& -4.3991( \pm 1.3665)^{1} \chi-3.8126( \pm 1.5876)^{2} \chi \\
& +2.8774( \pm 1.3718)^{3} \chi \tag{5}
\end{align*}
$$

$N=48, S E=0.5508, R^{2}=0.9157, R^{2} A=0.8942, F=40.703$

However, both these models (eqns 4 and 5) contain highly linearly correlated parameters (Table 7) and need to be examined for defect due to co-linearity. Several statistical techniques are available for examining the problem arose due to co-linearity. The chief among them being estimation of variance inflation factor (VIF), tolerance, eigenvalues and condition number. We have used all these four parameters for examining co-linearity in the nine- and the 10 -variable models mentioned above. We define these parameters as below.

The VIF is defined (12) as:

$$
\mathrm{VIF}=1 /\left(1-R_{i}^{2}\right)
$$

where, $R_{i}$ is the multiple correlation coefficient of the $i$ th independent variable on all of the other independent variables. In fact, a VIF is defined for each variable in the equation, and not the equation as a whole. So there must be as many VIFs as there are the number of correlating parameters in the proposed model. A VIF of 10 or more for large data set indicates a co-linearity problem. For small data sets, even VIFs of five or more can signify co-linearity. The variables with high VIFs are candidates for exclusion from the model. No higher limit is prescribed for VIF, however, the higher the value of VIF, the greater will be the problem due to co-linearity.

## Singh et al.

| Model number | Parameters used |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 1 | ${ }^{2} \chi^{v}$ |  |  |  |

Table 6: Results of variable selection for multiple regression analysis

Table 7: The values of parameters involved in models expressed by eqns 4 and 5

| Parameter | VIF | Tolerance | $\lambda_{i}$ | Condition number |
| :---: | :---: | :---: | :---: | :---: |
| (i) Equation 4 |  |  |  |  |
| W | 41.3772 | 0.0242 | 6.4117 | 1.00 |
| ${ }^{0} \chi$ | 4706.7683 | 0.0002 | 2.3594 | 2.72 |
| ${ }^{1} \chi$ | 951.6981 | 0.0011 | 0.1744 | 36.76 |
| ${ }^{3} \chi$ | 1531.1088 | 0.0007 | 0.0290 | 220.79 |
| ${ }^{2} \chi^{v}$ | 1452.6982 | 0.0007 | 0.0202 | 317.63 |
| ${ }^{1} \chi^{\text {shape }}$ | 1465.0880 | 0.0007 | 0.0030 | 2075.81 |
| ${ }^{3} \chi^{\text {shape }}$ | 24.7490 | 0.0404 | 0.0016 | 3981.25 |
| PNSA1 | 217.9133 | 0.0046 | 0.0004 | 15047.14 |
| PNSA2 | 199.5267 | 0.0050 | 0.0001 | 36222.58 |
| (ii) Equation 5 |  |  |  |  |
| W | 43.9542 | 0.0228 | 7.2788 | 1.00 |
| ${ }^{0} \chi$ | 6119.3205 | 0.0002 | 2.4892 | 2.92 |
| ${ }_{2}^{1} \chi$ | 972.9943 | 0.0010 | 0.1753 | 41.52 |
| ${ }^{2} \chi$ | 1365.9353 | 0.0007 | 0.0298 | 243.68 |
| ${ }_{3}^{3} \chi$ | 996.4784 | 0.0070 | 0.0193 | 376.39 |
| ${ }^{3} \chi^{v}$ | 680.0431 | 0.0015 | 0.0042 | 1717.41 |
| ${ }^{1} \chi \chi^{\text {shape }}$ | 1570.5717 | 0.0006 | 0.0017 | 4341.76 |
| ${ }^{3} \chi^{\text {shape }}$ | 38.5080 | 0.0260 | 0.0009 | 8355.81 |
| PNSA1 | 291.3843 | 0.0034 | 0.0006 | 12796.10 |
| PNSA2 | 276.2702 | 0.0036 | 0.0001 | 62056.74 |

All the variance inflation factor (VIF) are larger than 10, multi-collinearity is a problem.
All the tolerance values are very much smaller than 0.1 , multi-collinearity is a problem.
Some of the condition numbers are $>1000$, multi-collinearity is a problem.

The tolerance is just the denominator of VIF:

$$
\text { tolerance }=\left(1-R_{i}^{2}\right)
$$

The tolerance statistics is very effective in diagnosing multi-collinearity. Like VIF, the tolerance is also calculated for each of the independent variables present in the model.

The tolerance values range between 0 and 1. Paradoxically, high tolerance values indicate low multi-collinearity and low tolerance values indicate high multi-collinearity.

The eigenvalues of the correlation matrix is yet another technique for investigating multi-collinearity. The sum of the eigenvalues is equal to the number of independent variables. Eigenvalues near zero means that there is multi-collinearity in the proposed model.

The condition number is the largest eigenvalue divided by each corresponding eigenvalue. As the eigenvalues are real variance, the condition number is the ratio of variances. The condition number $>1000$ indicates the occurrence of severe multi-co-linearity problem, while condition numbers 100 and 1000 indicate a mild multi-collinearity problem.

All the aforementioned four parameters (VIF, tolerance, eigenvalues and condition number) are calculated (12) employing Ridge statistics and are used to resolve the problem due to multi-collinearity. For calculating these parameters we have used nCSS software ${ }^{\mathrm{e}}$.

We first discuss the abuse due to multi-collinearity in eqns 4 and 5 .
All the four parameters (VIF, tolerance, eigenvalues and condition number) for each of the descriptors involved in these equations are given in Table 7. We observe that massive co-linearity is present in both these models. For resolving this problem we have to delete parameters having highest VIF value in succession. When we did so, the nine-parametric model ended with two-variable model (eqn 6 , given below) free from the defect due to co-linearity. Likewise, the 10-parametric model ultimately yielded three-parametric model (eqn 7, given below); which is also free from co-linearity defect (Tables 8-10).

$$
\begin{align*}
\log \mathrm{T}(100)= & -3.6518+1.8371( \pm 0.1459)^{3} \chi^{v} \\
& +0.4719( \pm 0.3219)^{3} \chi^{\text {shape }}  \tag{6}\\
N=48, \mathrm{SE}=0.8135, & R^{2}=0.7789, R^{2} A=0.7691, F=79.284
\end{align*}
$$

Table 8: Correlation matrices for the parameters involved in eqns 6 and 7

|  | ${ }^{3} \chi^{v}$ |  | ${ }^{3} \chi^{\text {shape }}$ | $\log$ TA(100) |
| :---: | :---: | :---: | :---: | :---: |
| (i) Correlation matrix for eqn 6 |  |  |  |  |
| ${ }^{3} \chi^{v}$ | 1.0000 |  | -0.1256 | 0.8766 |
| ${ }^{3} \chi^{\text {shape }}$ | -0.1256 |  | 1.0000 | -0.0082 |
| $\operatorname{logTA}(100)$ | 0.8766 |  | -0.0082 | 1.0000 |
|  | ${ }^{3} \chi^{v}$ | ${ }^{3} \chi^{\text {shape }}$ | PNSA-2 | $\operatorname{logTA}(100)$ |
| (ii) Correlation matrix for eqn 7 |  |  |  |  |
| ${ }^{3} \chi^{v}$ | 1.0000 | -0.1256 | -0.2446 | 0.8766 |
| ${ }^{3} \chi^{\text {shape }}$ | -0.1256 | 1.0000 | -0.8128 | -0.0082 |
| PNSA-2 | -0.2447 | -0.8128 | 1.0000 | -0.2060 |
| $\operatorname{logTA}(100)$ | 0.8766 | -0.0082 | -0.2060 | 1.0000 |

Table 9: Variance inflation factors (VIF) and eigenvalues for the parameters involved in eqns 6 and 7

| Equations | Parameters | VIF | Eigenvalue $\left(\lambda_{i}\right)$ |
| :--- | :--- | :--- | :--- |
| 6 | ${ }^{3} \chi^{v}$ | 1.0160 | 1.1256 |
|  | ${ }^{v} \chi^{\text {shape }}$ | 1.0160 | 0.8744 |
| 7 | $\chi^{v} \chi^{v}$ | 1.5872 | 1.8219 |
|  | ${ }^{3} \chi^{\text {shape }}$ | 4.3965 | 1.0683 |
|  | PNSA-2 | 4.6026 | 0.1099 |

Table 10: ${ }^{\wedge}$-statistics

| Equations | Parameters | $\lambda_{i}$ | $1 / \lambda_{i}$ | $n$ | $\wedge$ |
| :--- | :--- | :--- | :--- | :--- | :--- |
| 6 | ${ }^{3} \chi^{v}$ | 1.1256 | 0.8884 | 2 | 1.0160 |
|  | ${ }^{3} \chi^{\text {shape }}$ | 0.8744 | 1.1436 |  |  |
| 7 | $\chi^{v} \chi^{v}$ | 1.8219 | 0.5489 | 3 | 3.2580 |
|  | ${ }^{3} \chi^{\text {shape }}$ | 1.0683 | 0.9360 |  |  |
|  | PNSA-2 | 0.1099 | 9.0992 |  |  |

$$
\begin{align*}
\log T A(100)= & -6.4990+2.1579( \pm 0.1659)^{3} \chi^{v} \\
& +2.1935( \pm 0.6091)^{3} \chi^{\text {shape }} \\
& +0.0146( \pm 0.0045) \text { PNSA- } 2 \tag{7}
\end{align*}
$$

$N=48, \mathrm{SE}=0.7399, R^{2}=0.8212, R^{2} A=0.8090, F=67.347$
Now, we have two two-parametric models (eqns 2 and 6) as well as two three-parametric models (eqns 3 and 7). The problem before us is to investigate which out of these pairs is the most appropriate model for modelling the activity. This problem can be resolved by examining the eqns 2 and 3 in the light of aforementioned four parameters (VIF, tolerance, eigenvalues and condition number). These values for the eqns 2 and 3, as recorded in Tables 11 and 12 indicate that co-linearity is present in both these models expressed by eqns 2 and 3 . The correlation matrix presented in Table 11 finally supports the occurrence of co-linearity in these models. However, no multi-collinearity exists in models expressed by eqns 5 and 6 (Tables 8-10). We, therefore, conclude that the models expressed by eqns 5 and 6 are the most appropriate and statistically significant models, free from defect due to co-linearity and that the three-variable model expressed by eqn 6 is the best for modelling, monitoring and estimating the activity. As stated earlier, we have examined the occurrence of outliers in the models expressed by eqns 5 and 6 and observed that compounds 4, 6, 9, 42, 46, 48 and 49 are outliers as the residues are two times larger than their standard deviations. The deletion of these seven compounds yielded the following models with improved statistics:

Table 11: Correlation matrices for eqns 2 and 3


Table 12: Variance inflation factor (VIF) values for the parameters involved in eqns 2 and 3

| (i) Equation 2 |  | (ii) Equation 3 |  |
| :---: | :---: | :---: | :---: |
| Parameter | VIF | Parameter | VIF |
| W | 13.0874 | W | 19.2255 |
| ${ }^{0} \chi^{v}$ | 13.0874 | ${ }^{3} \chi^{v}$ | 17.8443 |
|  |  | ${ }^{3} \chi^{\text {Shape }}$ | 4.2728 |

$$
\begin{aligned}
& \operatorname{logTA}(100)=-4.1811+1.9407( \pm 0.11250)^{3} \chi^{v} \\
&+0.6958( \pm 0.2670)^{3} \chi^{\text {shape }} \\
& N=41, \text { SE }=0.7009, R^{2}=0.8713, R^{2} A=0.8643, F=81.284 \\
& \operatorname{logTA}(100)=-6.3231+2.1830( \pm 0.1600)^{3} \chi^{v} \\
&+2.0277( \pm 0.5418)^{3} \chi^{\text {shape }} \\
&+0.0126( \pm 0.0042) \text { PNSA- } 2 \\
& N=41, \text { SE }=0.7434, R^{2}=0.8712, R^{2} A=0.8608, F=67.347
\end{aligned}
$$

We observed the statistics of both these models (eqns 8 and 9) are identical. Therefore, we conclude that two-parametric model (eqn 8) statistically better than the three-parametric model (eqn 9). These models show that mutagenicity is directly related to the size and shape of the compounds used in the present study.

Now we discuss Randic recommendations $(17,18)$ for resolving multi-collinearity in the models discussed above. Randic $(17,18)$ stated that 'the selection of descriptors to be used in structure-property-activity studies should not be delegated solely to the computers although the statistical criteria will continue to be useful for preliminary screening of descriptors taken from a large pool. Often in an automated selection of descriptors a descriptor will be discarded because it is highly correlated with another descriptor already selected. But what is important is not whether two descriptors parallel one another, i.e. duplicate much of the same structural information, but whether they in those parts are important for structure-property-activity correlations. If they differ in the domain which is important for the property-activity considered both descriptors should be retained; if they differ in parts that are not relevant for the correlation of the considered property-activity then one of them can be discarded. Hence, the residual of the correlation between two descriptors should be examined and kept or discarded depending on how well it can improve the correlation based on already selected descriptors'. Randic $(17,18)$ further stated that 'if a descriptor strongly correlates with another descriptor already used in a regression, such a descriptor in most studies should be discarded. For example, ${ }^{1} \chi$ and ${ }^{2} \chi,{ }^{1} \chi$ often strongly correlate and in many structure-property-activity studies ${ }^{2} \chi$ have been discarded. This is not theoretically justified and despite the widespread practice should be stopped. Although two highly correlated descriptors overall depict the same features of molecular structure, it is important to recognize that even highly interrelated descriptors differ in some other structural traits. The difference between them may be relatively small but nevertheless very important for structure-property regression'. Randic $(17,18)$ further argued that 'the criteria for inclusion or exclusion of descriptors should not be based on parallelism between descriptors even if overwhelming, but should be based on whether the part in which two descriptors disagree is or is not relevant for the characterization of the property'.

The criteria for inclusion or exclusion of descriptors should not be based on parallelism between descriptors even if overwhelming, but should be based on whether the part in which two descrip-
tors disagree is or is not relevant for the characterization of the property considered. If the part in which the second descriptor differs from the first, regardless of how small it is, is relevant for the property under consideration, then the descriptor should be included. Randic $(17,18)$ further stated that the selection of descriptors to be used in structure-property-activity studies should not be delegated solely to computers, although statistical criteria will continue to be useful for preliminary screening of descriptors taken from a large pool. Often in an automated selection of descriptors, a descriptor will be discarded because it is highly correlated with another descriptor already selected. But what is important is not whether two descriptors parallel one another, i.e. duplicates much of the same structural information, but whether they are complementary in those parts that are important for structure-propertyactivity correlations. Hence, the residual of the correlation between two descriptors should be examined and kept or discarded depending on how well it can improve the correlation based on already selected descriptors.

If we honour Randic's recommendations $(17,18)$ then all the models given in Table 6 (though in them correlated parameters exists) can be considered significant. Following Randic $(17,18)$ recommendations, the nine- and 10-parametric models (eqns 4 and 5) will be excellent as they account for $91 \%$ variation in mutagenic activity, e.g. IogTA100. Using these models we have calculated $\log T A 100$ and compared them with the observed $\operatorname{logTA100}$ values. Such a comparison is shown in Table 13 and illustrated in Figures 1 and 2 respectively. These results show that the calculated values of $\log$ TA100 are close to observed values of $\log T A 100$. Also that, the values of $R_{\text {pred }}^{2} \quad(0.9134$ and 0.9167 ) indicate that the nine- and 10 -parametric models have better predictive power. Both these models (eqns 4 and 5) can, therefore, be considered statistically significant on the basis of Randic $(17,18)$ recommendation as well as from the following observations: (i) the models are in accordance with the recommendations made by Randic $(17,18)$; (ii) there is consistent increase in $R^{2} A$ when we arrive at these equations and (iii) in both the equations all the correlating parameters have considerably larger values than their corresponding standard deviations.

In support of our aforementioned results, we have performed $\wedge$-statistics on the models expressed by eqns 4 and 5 (Table 10). The ${ }^{\wedge}$-statistics measure of the seriousness of collinearity in the model and is defined as:

$$
\begin{equation*}
\lambda=1 / n \sum_{i=1} 1 / \lambda_{i} \tag{10}
\end{equation*}
$$

Where $n$ is the number of descriptors in the model, and $\lambda_{i}$ are the eigenvalues of the correlation matrix of the descriptors. A value of $\wedge>5$ taken to indicate that collinearity problem exist in the model. The $\wedge$ for eqns 6 and $7<5$ indicating that both the models are free from the defect due to co-linearity.

Another empirical criteria for the presence/absence of multi-collinearity is given by the reciprocal of eigenvalues, i.e. eqn 10.

Table 13: Comparison of observed and calculated activity [ $\log T A(100)]$ using eqns 4 and 5

| Compound number | Observed | Model (4) |  | Model (5) |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Calculated | Res. | Calculated | Res. |
| 1 | 3.87 | 4.076 | -0.206 | 4.098 | -0.228 |
| 2 | 2.27 | 2.333 | -0.063 | 2.478 | -0.208 |
| 3 | 4.63 | 4.135 | 0.495 | 4.182 | 0.448 |
| 4 | 4.09 | 3.890 | 0.200 | 3.913 | 0.177 |
| 5 | 4.74 | 3.890 | 0.850 | 3.913 | 0.827 |
| 6 | 2.69 | 2.091 | 0.599 | 2.192 | 0.498 |
| 7 | 3.00 | 2.692 | 0.308 | 2.678 | 0.322 |
| 8 | 3.05 | 1.808 | 1.242 | 1.769 | 1.281 |
| 9 | 1.27 | 1.935 | -0.665 | 1.862 | -0.592 |
| 10 | 3.31 | 2.668 | 0.642 | 2.709 | 0.601 |
| 11 | 2.60 | 2.791 | -0.191 | 2.739 | -0.139 |
| 12 | 2.17 | 3.104 | -0.934 | 3.038 | -0.868 |
| 13 | 2.09 | 2.719 | -0.629 | 2.695 | -0.605 |
| 14 | 0.37 | -0.298 | 0.668 | -0.274 | 0.644 |
| 15 | 1.08 | 0.691 | 0.389 | 0.613 | 0.467 |
| 16 | 1.79 | 1.491 | 0.299 | 1.539 | 0.251 |
| 17 | 0.28 | -0.836 | 1.116 | -0.729 | 1.009 |
| 18 | 0.97 | 1.370 | -0.400 | 1.354 | -0.384 |
| 19 | -0.51 | -0.806 | 0.296 | -0.811 | 0.301 |
| 20 | 0.72 | 1.201 | -0.481 | 1.143 | -0.423 |
| 21 | -2.10 | -2.268 | 0.168 | -2.303 | 0.203 |
| 22 | -1.26 | -1.161 | -0.099 | -1.159 | -0.101 |
| 23 | -1.29 | -0.822 | -0.468 | -0.801 | -0.489 |
| 24 | -0.63 | -1.143 | 0.513 | -1.054 | 0.424 |
| 25 | -1.34 | -1.194 | -0.146 | -1.138 | -0.202 |
| 26 | -1.30 | -0.695 | -0.605 | -0.773 | -0.527 |
| 27 | -0.72 | -0.741 | 0.021 | -0.623 | -0.097 |
| 28 | 0.08 | 0.355 | -0.275 | 0.226 | -0.146 |
| 29 | 0.46 | 0.339 | 0.121 | 0.329 | 0.131 |
| 30 | 0.55 | 0.091 | 0.459 | 0.098 | 0.452 |
| 31 | 1.12 | 0.715 | 0.405 | 0.605 | 0.515 |
| 32 | 0.16 | 0.447 | -0.287 | 0.516 | -0.356 |
| 33 | 1.01 | 0.470 | 0.540 | 0.386 | 0.624 |
| 34 | 0.08 | 0.050 | 0.030 | 0.197 | -0.117 |
| 35 | -0.70 | 0.034 | -0.734 | 0.123 | -0.823 |
| 36 | 0.86 | 1.174 | -0.314 | 1.300 | -0.440 |
| 37 | 0.91 | 0.636 | 0.274 | 0.861 | 0.049 |
| 38 | 1.12 | 1.614 | -0.494 | 1.456 | -0.336 |
| 39 | 2.46 | 2.576 | -0.116 | 2.450 | 0.010 |
| 40 | 2.87 | 2.922 | -0.052 | 2.848 | 0.022 |
| 41 | 3.18 | 3.377 | -0.197 | 3.375 | -0.195 |
| 42 | -0.70 | -0.736 | 0.036 | -0.783 | 0.083 |
| 43 | -1.05 | -0.222 | -0.828 | -0.387 | -0.663 |
| 44 | -0.30 | -0.367 | 0.067 | -0.498 | 0.198 |
| 45 | -1.00 | -0.424 | -0.576 | -0.542 | -0.458 |
| 46 | -0.30 | -0.455 | 0.155 | -0.528 | 0.228 |
| 47 | 0.26 | 1.096 | -0.836 | 1.282 | -1.022 |
| 48 | 2.21 | 2.509 | -0.299 | 2.551 | -0.341 |

Res. $=$ difference between observe and calculated activity $[\log T A(100)]$

If this sum is greater than five times the number of predictor variable, then the collinearity is present. In our case this sums are 2.0320 and 10.5841 , respectively, for eqns 6 and 7 . Which are much smaller than five times the number of descriptors, i.e. 10 and 15 respectively.

Once again these results indicate absence of collinearity in models expressed by eqns 4 and 5 respectively. Finally, we have also calculated condition number $k$ using following expression:

$$
k=\frac{\text { Maximum eigenvalue of the correlation matrix }}{\sqrt{\text { Minimum eigenvalue of the correlation matrix }}}
$$

$$
\begin{equation*}
k=\sqrt{\left(\lambda_{i} / \lambda_{p}\right)} \tag{11}
\end{equation*}
$$

It is interesting to mention that condition number $k$ will always be $>10$. A larger condition number indicates evidences of strong collinearity. The co-linearity problem is massive if the condition number exceeds 15 (which means that $\lambda_{i}$ is more than 25 times $\lambda_{p}$ ). In our case $k$ is found to be 1.1346 and 4.0716 , respectively, for models expressed by eqns 6 and 7 respectively. Hence, $k$ also shows that the proposed models are free from co-linearity problems.

Predictive ability was evaluated by the LOO cross-validation procedure (12). This method systematically removes one data point at a time. A model is constructed on the basis of this reduced data set and is subsequently used to predict the removed sample. This procedure was repeated for all points until a complete set of predicted values was obtained. The following criteria were used for the quality of predictive ability: predictive residual sum of squares (PRESS), sum of the squares of the response values (SSY), squared correlation coefficient of prediction $\left(Q^{2}\right)$, uncertainty of prediction ( $\left.S_{\text {press }}\right)$ and standard error of prediction (SDEP). These criteria were calculated as follows:

$$
\begin{aligned}
\text { PRESS } & =\sum\left(\log A_{\text {pred }}-\log A_{\text {obs }}\right)^{2} \\
\text { SSY } & =\sum\left(\log A_{\text {obs }}-\log A_{\text {mean }}\right)^{2} \\
Q^{2} & =(\text { SSY }- \text { PRESS }) / \text { SSY } \\
S_{\text {press }} & =(\text { PRESS } / n-k-1)^{1 / 2} \\
\text { SDEP } & =(\text { PRESS } / n)^{1 / 2}
\end{aligned}
$$

Here, $n$ is the number of compounds, $k$ is the number of variables in the model, $A_{\text {pred }}$ is predicted activity, $A_{\text {obs }}$ is the observed activity and $A_{\text {mean }}$ is the mean activity. The calculated values of these cross-validated parameters for the proposed models are presented in Table 14, which show that for all the models PRESS is smaller than SSY indicating that models predict better than chance and can be considered statistically significant. The ratio PRESS/SSY, smaller than 0.4 indicates that the models are reasonable OSAR models. This ratio smaller than 0.1 indicates that the models are excellent. The use of SDEP is more directly related to the uncertainty of prediction.

It will be interesting to compare our results with those reported by Gramatica and co-workers (11). In their report Gramatica and coworkers (11) two models for the full set of 48 compounds. Both these models were two-variable models containing: (i) CICI, PW2 and (ii) LUMO, MR as the correlating parameters. As the parameters used by Gramatica and co-workers (11) are quite different from

## Singh et al.




Figure 1: Correlation of observed and calculated (estimated) activity (logTA100) using model (eqn 4).

Figure 2: Correlation of observed and calculated (estimated) activity (logTA100) using model (eqn 5).

Table 14: Cross-validated parameters for the proposed models

| Model <br> (equation) | PRESS | SSY | PRESS/ <br> SSY | $0^{2}$ | $S_{\text {press }}$ | SDEP |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 2 | 57.0593 | 114.1185 | 0.5000 | 0.4999 | 1.0612 | 1.0903 |
| 3 | 39.1897 | 117.5692 | 0.3333 | 0.6666 | 0.9438 | 0.0036 |
| 4 | 14.3178 | 128.8598 | 0.1111 | 0.8888 | 0.6138 | 0.5462 |
| 5 | 12.9318 | 128.3176 | 0.1008 | 0.8992 | 0.5912 | 0.5190 |
| 6 | 54.9438 | 109.8877 | 0.4999 | 0.5000 | 1.1049 | 1.0698 |
| 7 | 38.6148 | 115.8487 | 0.3333 | 0.6666 | 0.9368 | 0.8969 |

those used by us, the exact comparison is not possible. However, based on the regression statistics the best model proposed by us (eqn 5) is better than the model (ii) proposed by Gramatica and coworkers (11), while it is inferior to the model (i) of Gramatica and co-workers (11). Furthermore, for the reduced set of 41 compounds the model proposed by us (eqn 8 ) is the best model for modelling the activity. This is far superior to Gramatica and co-workers (11) model as the reduced set of compounds used by them contain 31
compounds only. This comparison demonstrates the validity of particular combinations of molecular descriptors vis-à-vis a particular combination of structural information in the studied response prediction. The most relevant used descriptor is CICl , an information content index based on the calculation of equivalence classes in the molecular graph of the compounds used. This descriptor is positively related to mutagenicity and gives information on molecular size and increases with the number of rings present in the compounds and nitro-group in each series of congeners. Another important descriptor is PW2 related to the shape of the molecule and is also directly correlated to mutagenicity. This index is called path/ walk-2 Randic shape index. When compared to LUMO and MR, ${ }^{3} \chi^{v}$ and ${ }^{3} \chi^{\text {shape }}$ indices are better descriptors to model the activity. Same is the case with the descriptor PNA-2.

## Conclusions

From the Results and discussions made above we conclude that the methodology used by us to the estimation of the mutagenic activity

## Mutagenicity of Nitrated Polycyclic Aromatic Hydrocarbons

is quite simple as it requires only two or three correlating parameters devoid of any multi-collinearity for estimating the mutagenicity. Also, that following recommendations of Randic $(22,23)$ even the nine- and 10-parametric models, containing highly correlated parameters can be considered statistically significant.

## Acknowledgments

One of the authors, Shalini Singh express her thanks to the Department of Science \& Technology, Government of India, New Delhi, for awarding DST project SR/WOS-A/CS/61/2004 under Woman Scientists scheme, and to Principal for his personal interest and providing facility to carry out this work.

## References

1. Hansch C., Leo A. (1995) Exploring OSAR Fundamentals and Applications in Chemistry and Biology, ACS Professional Reference Book. American Chemical Society, Washington, DC, 1995, pp. 557.
2. Karelson M. (2000) Molecular Descriptors in QSAR/QSPR. J Wiley \& Sons, New York (NY), pp. 430.
3. Diudea M.V., Florescn M.S., Khadikar P.V. (2006) Molecular Topology and its Applications. Bucarest: EFICON Press.
4. Hansch C., Telzer B.R., Zhang L. (1995) Comparative OSAR in toxicology: example from teratology and cancer chemotherapy of aniline mustard. Crit Rev Toxicol;25:67-89.
5. Branes H. (2004) Mammalian toxicology - Property Prediction and OSAR Technique. Kent: Peter Fisk Associates, Whitestable
6. Bradbury S. (1995) Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research. Taxicol Lett;79:229-237.
7. Duchowicz P.R., Castro E.A., Toropov A.A., Benfenati E. (2006) Application of flexible molecular descriptors in the QSPR/QSAR study of hetero-cyclic drugs. Top Heterocycl Chem;3:1-18.
8. Fereira M.M.C. (2001) Polycyclic aromatic hydrocarbons: a OSAR study. Chemosphere;44:125-146.
9. Vracko M. (2006) QSAR approach in study of mutagenicity of aromatic and heterocyclic amines. Top Heterocycl Chem;4:85-106.
10. Gallegos A., Robert G., Gerones X., Carbo-Dorca R. (2001) Struc-ture-toxicity correlationship of polycyclic aromatic hydrocarbons using molecular quantum similarity. J Comput Aided Mol Des;15:67-80.
11. Gramatica P., Pilutti P., Papa E. (2007) Approaches for externally validated OSAR modeling of nitrated polycyclic hydrocarbon mutagenicity. SAR-OSAR Environ Res;18:169-178.
12. Chaterjee S., Hadi A.S., Price B. (2000) Regression Analysis by Examples, 3rd edn. New York: Wiley.
13. Singh J., Singh S., Meer S., Agrawal V.K., Khadikar P.V., Balaban A.T. (2006) QSAR correlations of half-wave reduction potentials of cata-condensed benzenoid hydrocarbons. Arkivoc;14:103-118.
14. Singh J., Lakhwani M., Khadikar P.V., Balaban A.T., Clare B.W., Supuran C.T. (2006) QSAR study on the inhibition of the human carbonic anhydrase cystolic isozyme VII. Rev Roum De Chim;51: 691-701.
15. Khadikar P.V., Clare B.W., Balaban A.T., Supuran C.T., Agrawal V.K., Singh J., Joshi A.K., Lakhwani M. (2006) OSAR modeling of carbonic anhydrase-I, -II, and -IV inhibitory activities: relative correlation potential of six topological indices. Rev Roum De Chim;51:703-717.
16. Tute M.S. (1971) History and objectives of quantitative drug design. In Quantitative Drug Design-Vol.4.: (Ramsden C.A., ed), Pergamon Press, Oxford (UK). pp. 1-31.
17. Randic M. (1998) On the characterisation of molecular attributes. Acta Chem Slov;45:239.
18. Randic M. (1997) On the characterisation of chemical structure. J Chem Inf Comput Sci;37:672.

## Notes

${ }^{\text {a }}$ Todeschini R.; Consonni V.; Mauri A.; Pavan M. (2005) dragon, version 5.3, Milan, Italy: Talete srl.
${ }^{\text {b }}$ Karelson, M. Chemaxon (http://www.chemaxon.com) softwares for the calculation of topological indices.
${ }^{\text {}}$ нYPERCHEm, Release 7.03 for Windows (2002) Aorida, USA: Hypercube, Inc.
${ }^{d}$ ACD-LAB software for calculating the referred physicochemical parameters; сНем SKETCH 3.0. Available at: http://www.acdlabs.com.
${ }^{e}$ NcSs. Available at: http://www.ncss.com.

## Appendix-A

| Wiener index: | W |
| :--- | :--- |
| Randic index (order 0): | ${ }^{0} \chi$ |
| Randic index (order 1): | $1^{\chi}$ |
| Randic index (order 2): | ${ }^{2} \chi$ |
| Randic index (order 3): | ${ }^{3}$ |
| Kier\&Hall index (order 0): | $\chi^{v}$ |
| Kier\&Hall index (order 1): | $\chi^{1} \chi^{v}$ |
| Kier\&Hall index (order 2): | $\chi^{v} \chi^{v}$ |
| Kier\&Hall index (order 3): | $\chi^{v}$ |
| Kier shape index (order 1): | $\chi^{\text {shape }}$ |
| Kier shape index (order 2): | $\chi^{\text {shape }}$ |
| Kier shape index (order 3): | $\chi^{\text {shape }}$ |
| Total molecular surface area [Empirical PC]: | TMSA |
| Partial positive surface area [Empirical PC]: | PPSA1 |
| Total charge weighted PPSA [Empirical PC]: | PPSA2 |
| Atomic charge weighted PPSA [Empirical PC]: | PPSA3 |
| Partial negative surface area [Empirical PC]: | PNSA1 |
| Total charge weighted PNSA [Empirical PC]: | PNSA2 |
| Atomic charge weighted PNSA [Empirical PC]: | PNSA3 |
| Difference in CPSAs (PPSA1-PNSA1) [Empirical PC]: | DPSA1 |
| Difference in CPSAs (PPSA2-PNSA2) [Empirical PC]: | DPSA2 |
| Difference in CPSAs (PPSA3-PNSA3) [Empirical PC]: | DPSA3 |
| Fractional PPSA (PPSA-1/TMSA) [Empirical PC]: | FPSA1 |
| Fractional PPSA (PPSA-2/TMSA) [Empirical PC]: | FPSA2 |
| Fractional PPSA (PPSA-3/TMSA) [Empirical PC]: | FPSA3 |
| Fractional PNSA (PNSA-1/TMSA) [Empirical PC]: | FNSA1 |
| Fractional PNSA (PNSA-2/TMSA) [Empirical PC]: | FNSA2 |
| Fractional PNSA (PNSA-3/TMSA) [Empirical PC]: | FNSA3 |

