Int. J. Mol. Sci. **2014**, 15, 1-x manuscripts; doi:10.3390/ijms150x000x

**OPEN ACCESS**

International Journal of  
Molecular Sciences

ISSN 1422-0067  
www.mdpi.com/journal/ijms

*Article*

**Title of the Paper (M\_Title)**

**Subhash C. Basak 1,\*, Subhabrata Majumdar 2**

1 *Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*; E-Mail: sbasak@nrri.umn.edu

2 *School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA*; E-Mail: majum010@umn.edu

**\*** Author to whom correspondence should be addressed; E-Mail: sbasak@nrri.umn.edu;   
Tel.: +1 (218)720-4230; Fax: +1 (218)720-4328.

*Received: / Accepted: / Published:*

**Abstract:** This is the abstract section. One paragraph only (Maximum 200 words).

**Keywords:** Mutagenicity; interrelated two-way clustering; topological indices; ridge regression; hierarchical QSAR

**1. Introduction**

Main text paragraph.

Main text paragraph.

**2. Results and Discussion**

Main text paragraph.

Main text paragraph.

|  |  |
| --- | --- |
| [add an equation here; use MS Word or MathType equation function] | (1) |

(Note: all equations should be completed within a two column table with one line, centered, no boarders, as example see above).

Main text paragraph.

Main text paragraph.

*2.1. This is Subsection Heading*

Main text paragraph.

Main text paragraph.

2.1.1. This is Subsection Heading

Main text paragraph.

Main text paragraph.

**Table 1.** Comparison of model performances for diverse and congeneric datasets

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset used** | **Predictive model** | **Type of predictor used** | **No. of predictors** | **Correct classification %** | **Sensitivity** | **Specificity** |
| **508 compound diverse dataset** | Ridge regression  [hawk] | TS+TC | 298 | 76.97 | 83.98 | 69.84 |
| TS+TC+3D+QC | 307 | 77.17 | 84.38 | 69.84 |
| ITC + Ridge regrerssion [itc508] | TS+TC+AP | 203 | 78.35 | 84.38 | 72.22 |
| **95 amines congeneric dataset** | Ridge regression | TS | 97 | 83.16 | 75.47 | 88.42 |
| TS+TC | 259 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D | 269 | 84.21 | 77.36 | 92.86 |
| TS+TC+3D+QC | 275 | 84.21 | 77.36 | 92.86 |
| ITC + Ridge regrerssion | TS | 97 | 88.42 | 90.57 | 85.71 |
| TS+TC | 259 | 86.32 | 88.68 | 83.33 |
| TS+TC+3D | 269 | 88.42 | 92.45 | 83.33 |
| TS+TC+3D+QC | 275 | 85.26 | 88.68 | 80.95 |

**Figure 1.** (**a**) Add a descriptive label of the figure here. (**b**) Add a descriptive label of the figure here. (**c**) Add a descriptive label of the figure here.

[add the figure here]

**3. Experimental Section**

*3.1. Data*

The dataset used in our analysis is due to Debnath *et al* [debnath] and concerns mutagenic activities of 95 aromatic and heteroaromatic amines in *S. typhimurium* TA98+S9 microsomal preparations. There are 275 descriptors for each compound: among them 97 topostructural, 162 topochemical, 10 3-dimensional and 6 quantum-chemical. Number of revertants per nmol of test culture in log scale (log *R*) is the original response variable. For binary classification we take the 0/1 indicator log *R* being > 0 or < 0 as the response variable in our analysis. At the original scale, this amounts to no. of revertants/nmol being greater/less than 1, respectively.

The results obtained are compared with previous studies on a second, and more diverse, dataset records Ames mutagenicity of 508 chemical compounds (256 mutagens and 252 non-mutagens), and is taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [crc]. The set of descriptors for these compounds includes the above type of descriptors, as well as a large number of Atom-pair (AP) descriptors.

*3.2. Methods*

3.2.1. Variable selection

Because the number of samples (*n* = 95) is smaller than the number of variables (*p* = 275), regression-based methods of variable selection like stepwise forward and backward selection are not applicable in our scenario. Here we use the Interrelated Two-way Clustering (ITC) algorithm to do the variable selection. This algorithm takes in a number of predictor groups, which are pre-determined or obtained by some known unsupervised clustering method (like K-means). It then hypothesizes that if we include only the important predictors and classify samples independently for each predictor groups, the classification should be identical. Keeping this in mind, an iterative procedure is used to eliminate predictors until the classifications based on different predictor groups achieve a certain level of similarity.

A detailed description of the algorithm can be found in the original paper [itc], and its implementation in QSAR was done by Majumdar *et al* [itc508].

3.2.2. Hierarchical QSAR and predictive models

We take a hierarchical approach to build the predictive model. Starting from topostructural variables, we keep on including topochemical, 3-dimensional and quantum-chemical variables and check predictive performance of all the models. To tackle high collinearity among different predictors, we use ridge regression to build our predictive models. Given *n* samples and *p* variables, the *n* × *p* data matrix of predictors ***X*** and *n* × *1* vector of 0/1 responses ***Y***, the vector of coefficients obtained by ridge regression is defined as:

|  |  |
| --- | --- |
| ***b*** *=* (***X****’****X*** *+ k****I***)-1***Y*** | (1) |

Where *k* > 0 is the ridge constant, chosen by cross-validation [ridge-book].

While assessing the predictive performance of a model through cross validation, it is essential to not do the variable selection beforehand and the build the model, because that uses information from the test set of compounds in the variable selection step. This results in a synthetic increase of the predictive performance of the model. For this reason, while we did use leave-one-out cross-validation to obtain prediction accuracy, we did both the selection of variables and building ridge regression models for every iteration of the cross-validation procedure.

**4. Conclusions**

Main text paragraph.

Main text paragraph.

**Acknowledgments**

Main text paragraph.

**Author Contributions**

Main text paragraph.

**Conflicts of Interest**

The authors confirm that this article has no conflict of interest.

**References and Notes**

debnath. Debnath, A. K.; Debnath, G.; Shusterman, A. J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Muagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.* **1992**, *19*, 37-52.

crc. Soderman, J.V. *CRC Handbook of Identified Carcinogens and Noncarcinogens: Carcinogenicity-Mutagenicity Database*, CRC Press: Boca Raton, FL, **1982**.

itc. Tang, C.; Zhang, L.; Zhang, A.; Ramanathan, M. In: *Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis*, Proceedings of BIBE 2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, November 4-5, 2001; Bilof, R.; Palagi, L., Eds.; IEEE Computer Society: Los Alamitos, CA, **2001**; pp. 41-48.

itc508. Majumdar, S.; Basak, S. C.; Grunwald, G. D. Adapting Interrelated Two-Way Clustering Method for Quantitative Structure-Activity Relationship (QSAR) Modeling of Mutagenicity/Non-Mutagenicity of a Diverse Set of Chemicals. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 463-471.

ridge-book. Yang, X.; Su, X.G. *Linear Regression Analysis- Theory and Computing*, World Scientific: Singapore, **2009**.

hawk. Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 37-44.

2. Author 1, A.; Author 2, B. Title of the chapter. In *Book Title*, 2nd ed.; Editor 1, Editor 2, Eds.; Publisher: Publisher Location, Country, 2007; Volume 3, pp. 154–196.

3. Author 1, A.; Author 2, B. *Book Title*, 3rd ed.; Publisher: Publisher Location, Country, 2008;   
pp. 154–196.

4. Author 1, A.B.; Author 2, C. Title of Unpublished Work. Journal Abbreviation, phrase indicating stage of publication.

5. Author 1, A.B.; Author 2, C.D.; Author 3, E.F. Title of Presentation. In *Title of the Collected Work* (if available), Proceedings of the Name of the Conference, Location of Conference, Country, Date of Conference; Editor 1, Editor 2, Eds. (if available); Publisher: City, Country, Year (if available); Abstract Number (optional), Pagination (optional).

6. Author 1, A.B. Title of Thesis. Level of Thesis, Degree-Granting University, Location of University, Date of Completion.

7. Author 1, A.B.; Author 2, C.D. Title of the article. *Abbreviated Journal Name*, Year, Volume, (Page range), doi or other identification number. Available online: http://URL (accessed on Day Month Year).

8. Title of Site. Available online: http://URL (accessed on Day Month Year).

Reference list: We recommend the use of reference management software to prepare the references list (e.g., Endnote, http://www.mdpi.com/files/word-templates/MDPI.ens).

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).

**Appendix**

Table 2. Symbols, definitions and classification of topological indices

|  |  |  |  |
| --- | --- | --- | --- |
|  | Topostructural (TS) | In 508 compound data? | In 95  Amine data? |
| *IWD* | Information index for the magnitudes of distances between all possible pairs of vertices of a graph | **✓** | **✓** |
| *IWD* | Mean information index for the magnitude of distance | **✓** | **✓** |
| *W* | Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph | **✓** | **✓** |
| *ID* | Degree complexity | **✓** | **✓** |
| *HV* | Graph vertex complexity | **✓** | **✓** |
| *HD* | Graph distance complexity | **✓** | **✓** |
| *IC* | Information content of the distance matrix partitioned by frequency of occurrences of distance *h* | **✓** | **✓** |
| *M1* | A Zagreb group parameter = sum of square of degree over all vertices | **✓** | **✓** |
| *M2* | A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices | **✓** | **✓** |
| *hχ* | Path connectivity index of order *h* = 0-10 | **✓** | **✓** |
| *hχC* | Cluster connectivity index of order *h* = 3-5 | **✓** | **✓** |
| *6χC* | Cluster connectivity index of order *6* | **✓** | **✗** |
| *hχPC* | Path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *hχCh* | Chain connectivity index of order *h* = 3, 4, 7, 8 | **✓** | **✗** |
| *hχCh* | Chain connectivity index of order *h* = 5, 6, 9, 10 | **✓** | **✓** |
| *Ph* | Number of paths of length *h* = 0-10 | **✓** | **✓** |
| *J* | Balaban’s *J* index based on topological distance | **✓** | **✓** |
| *nrings* | Number of rings in a graph | **✓** | **✓** |
| *ncirc* | Number of circuits in a graph | **✓** | **✓** |
| *DN2Sy* | Triplet index from distance matrix, square of graph order, and distance sum; operation *y* = 1-4 | **✓** | **✓** |
| *DN21y* | Triplet index from distance matrix, square of graph order, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| *DN215* | Triplet index from distance matrix, square of graph order, and number 1; operation *5* | **✓** | **✗** |
| *AS1y* | Triplet index from adjacency matrix, distance sum, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| *DS1y* | Triplet index from distance matrix, distance sum, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| *ASNy* | Triplet index from adjacency matrix, distance sum, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| *DSNy* | Triplet index from distance matrix, distance sum, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| *DN2Ny* | Triplet index from distance matrix, square of graph order, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| *ANSy* | Triplet index from adjacency matrix, graph order, and distance sum; operation *y* = 1-5 | **✓** | **✓** |
| *AN1y* | Triplet index from adjacency matrix, graph order, and number 1; operation *y* = 1-5 | **✓** | **✓** |
| *ANNy* | Triplet index from adjacency matrix, graph order, and graph order again; operation *y* = 1-4 | **✓** | **✓** |
| *ANN5* | Triplet index from adjacency matrix, graph order, and graph order again; operation *5* | **✓** | **✗** |
| *ASVy* | Triplet index from adjacency matrix, distance sum, and vertex degree; operation *y* = 1-5 | **✓** | **✓** |
| *DSVy* | Triplet index from distance matrix, distance sum, and vertex degree; operation *y* = 1-2 | **✓** | **✓** |
| *DSVy* | Triplet index from distance matrix, distance sum, and vertex degree; operation *y* = 3-5 | **✓** | **✗** |
| *ANVy* | Triplet index from adjacency matrix, graph order, and vertex degree; operation *y* = 1-5 | **✓** | **✓** |
| *kp0* | Kappa zero | **✓** | **✓** |
| *kp1-kp3* | Kappa simple indices | **✓** | **✓** |
|  | **Topochemical (TC)** | **In 95 amine data?** | **In 508 compound data?** |
| *O* | Order of neighborhood when *ICr* reaches its maximum value for the hydrogen-filled graph | **✓** | **✓** |
| *Oorb* | Order of neighborhood when *ICr* reaches its maximum value for the hydrogen-suppressed graph | **✓** | **✓** |
| *IORB* | Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices | **✓** | **✓** |
| *ICr* | Mean information content or complexity of a graph based on the *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| *SICr* | Structural information content for *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| *CICr* | Complementary information content for *r*th (*r* = 0-6) order neighborhood of vertices in a hydrogen-filled graph | **✓** | **✓** |
| *hχb* | Bond path connectivity index of order *h* = 0-6 | **✓** | **✓** |
| *hχbC* | Bond cluster connectivity index of order *h* = 3, 5 | **✓** | **✓** |
| *hχbC* | Bond cluster connectivity index of order *h* = 4, 6 | **✓** | **✗** |
| *hχbCh* | Bond chain connectivity index of order *h* = 3, 4 | **✓** | **✗** |
| *hχbCh* | Bond chain connectivity index of order *h* = 5, 6 | **✓** | **✓** |
| *hχbPC* | Bond path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *hχv* | Valence path connectivity index of order *h* = 0-10 | **✓** | **✓** |
| *hχvC* | Valence cluster connectivity index of order *h* = 3, 5 | **✓** | **✓** |
| *hχvC* | Valence cluster connectivity index of order *h* = 4, 6 | **✓** | **✗** |
| *hχvCh* | Valence chain connectivity index of order *h* = 3, 4, 7, 8 | **✓** | **✗** |
| *hχvCh* | Valence chain connectivity index of order *h* = 5, 6, 9, 10 | **✓** | **✓** |
| *hχvPC* | Valence path-cluster connectivity index of order *h* = 4-6 | **✓** | **✓** |
| *JB* | Balaban’s*J* index based on bond types | **✓** | **✓** |
| *JX* | Balaban’s*J* index based on relative electronegativities | **✓** | **✓** |
| *JY* | Balaban’s*J* index based on relative covalent radii | **✓** | **✓** |
| *AZVy* | Triplet index from adjacency matrix, atomic number, and vertex degree; operation *y* = 1-5 | **✓** | **✓** |
| *AZSy* | Triplet index from adjacency matrix, atomic number, and distance sum; operation *y* = 1-5 | **✓** | **✓** |
| *ASZy* | Triplet index from adjacency matrix, distance sum, and atomic number; operation *y* = 1-5 | **✓** | **✓** |
| *AZNy* | Triplet index from adjacency matrix, atomic number, and graph order; operation *y* = 1-5 | **✓** | **✓** |
| *ANZy* | Triplet index from adjacency matrix, graph order, and atomic number; operation *y* = 1-5 | **✓** | **✓** |
| *DSZy* | Triplet index from distance matrix, distance sum, and atomic number; operation *y* = 1,2 | **✓** | **✓** |
| *DSZy* | Triplet index from distance matrix, distance sum, and atomic number; operation *y* = 3-5 | **✓** | **✗** |
| *DN2Zy* | Triplet index from distance matrix, square of graph order, and atomic number; operation *y* = 1-4 | **✓** | **✓** |
| *DN2Zy* | Triplet index from distance matrix, square of graph order, and atomic number; operation *5* | **✓** | **✗** |
| *nvx* | Number of non-hydrogen atoms in a molecule | **✓** | **✓** |
| *nelem* | Number of elements in a molecule | **✓** | **✓** |
| *fw* | Molecular weight | **✓** | **✓** |
| *si* | Shannon information index | **✓** | **✓** |
| *totop* | Total Topological Index *t* | **✓** | **✓** |
| *sumI* | Sum of the intrinsic state values *I* | **✓** | **✓** |
| *sumdelI* | Sum of delta-*I* values | **✓** | **✓** |
| *tets2* | Total topological state index based on electrotopological state indices | **✓** | **✓** |
| *phia* | Flexibility index (*kp*1\* *kp*2/*nvx*) | **✓** | **✓** |
| *Idcbar* | Bonchev-Trinajstić information index | **✓** | **✓** |
| *IdC* | Bonchev-Trinajstić information index | **✓** | **✓** |
| *Wp* | Wienerp | **✓** | ✓ |
| *Pf* | Plattf | **✓** | ✓ |
| *Wt* | Total Wiener number | **✓** | ✓ |
| *knotp* | Difference of chi-cluster-3 and path/cluster-4 | **✓** | ✓ |
| *knotpv* | Valence difference of chi-cluster-3 and path/cluster-4 | **✓** | ✓ |
| *nclass* | Number of classes of topologically (symmetry) equivalent graph vertices | **✓** | ✓ |
| *NumHBd* | Number of hydrogen bond donors | **✓** | ✓ |
| *NumHBa* | Number of hydrogen bond acceptors | **✓** | ✓ |
| *SHCsats* | E-State of C *sp3* bonded to other saturated C atoms | **✓** | ✓ |
| *SHCsatu* | E-State of C *sp3* bonded to unsaturated C atoms | **✓** | ✓ |
| *SHvin* | E-State of C atoms in the vinyl group, *=CH-* | **✓** | ✗ |
| *SHtvin* | E-State of C atoms in the terminal vinyl group, *=CH2* | **✓** | ✗ |
| *SHavin* | E-State of C atoms in the vinyl group, *=CH-*, bonded to an aromatic C | **✓** | ✗ |
| *SHarom* | E-State of C *sp2* which are part of an aromatic system | **✓** | ✓ |
| *SHHBd* | Hydrogen bond donor index, sum of Hydrogen E-State values for *–OH*, *=NH*, -*NH2*, *-NH-,-SH*, and *#CH* | **✓** | ✓ |
| *SHwHBd* | Weak hydrogen bond donor index, sum of *C-H* Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded | **✓** | ✗ |
| *SHHBa* | Hydrogen bond acceptor index, sum of the *E*-State values for *–OH*, *=NH*, *-NH2*, -*NH-*, *>N*, *-O-*, *-S-*, along with –F and –Cl | **✓** | ✓ |
| *Qv* | General Polarity descriptor | **✓** | ✓ |
| *NHBint2* | Count of potential internal hydrogen bonders (*y* = 2) | **✓** | ✗ |
| *NHBinty* | Count of potential internal hydrogen bonders (*y* = 3-10) | **✓** | ✓ |
| *SHBinty* | E-State descriptors of potential internal hydrogen bond strength (*y =2, 5-10*) | **✓** | ✗ |
| *SHBinty* | E-State descriptors of potential internal hydrogen bond strength (*y =3, 4*) | **✓** | ✓ |
| *ka1-ka3* | Kappa alpha indices | **✓** | ✓ |
|  | Electrotopological State index values for atom types:  *SHdNH, SHsSH, HssNH, SHtCH, SHCHnX, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH ,SsssB, SssssBm, SdCH2, StCH, SddC, StsC, SdssC, SaasC, SsNH3p, SssNH2p, SdNH, StN, SsssNHp, SdsN, SsssN, SaasN, SssssNp, SaaO, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SaaS, SdssS, SddssS, SssssssS, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb* | **✓** | **✗** |
|  | Electrotopological State index values for atom types:  *SHsOH, SHsNH2, SHssNH, SHother, Hmax, Gmax, Hmin, Gmin, SsCH3, SssCH2, SdsCH, SaaCH, SsssCH, SaaaC, SssssC, SsNH2, SssNH, SaaNH, SaaN, SddsN, SsOH, SdO, SssO, SsF, SssS, SsCl, SsBr,* | **✓** | **✓** |
|  | Electrotopological State index values for atom types:  *SHssNH* | **✗** | **✓** |
|  | **Geometrical (3-D)** | **In 95 amine data?** | **In 508 compound data?** |
| *3DW* | 3D Wiener number based on the hydrogen-suppressed geometric distance matrix | **✓** | **✓** |
| *3DW H* | 3D Wiener number based on the hydrogen-filled geometric distance matrix | **✓** | **✓** |
| *VW* | Van der Waal’s volume | **✓** | **✓** |
|  | Quantum Chemical (QC) | **In 95 amine data?** | **In 508 compound data?** |
| *EHOMO* | Energy of the highest occupied molecular orbital | **✓** | **✓** |
| *EHOMO-1* | Energy of the second highest occupied molecular | **✓** | **✓** |
| *ELUMO* | Energy of the lowest unoccupied molecular orbital | **✓** | **✓** |
| *ELUMO+1* | Energy of the second lowest unoccupied molecular orbital | **✓** | **✓** |
| *ΔHf* | Heat of formation | **✓** | **✓** |
| *μ* | Dipole moment | **✓** | **✓** |