Adapting Interrelated Two-Way Clustering Method for Quantitative Structure-Activity Relationship (QSAR) Modeling of Mutagenicity/Non-Mutagenicity of a Diverse Set of Chemicals

Subhabrata Majumdar[1], Subhash C. Basak[*,2] and Gregory D. Grunwald[2]

[1]*School of Statistics, University of Minnesota- Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA*

[2]*Natural Resources Research Institute, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*

Abstract: Interrelated Two-way Clustering (ITC) is an unsupervised clustering method developed to divide samples into two groups in gene expression data obtained through microarrays, selecting important genes simultaneously in the process. This has been found to be a better approach than conventional clustering methods like K-means or self-organizing map for the scenarios when number of samples is much smaller than number of variables ($n<<p$). In this paper we used the ITC approach for classification of a diverse set of 508 chemicals regarding mutagenicity. A large number of topological indices (TIs), 3-dimensional, and quantum chemical descriptors, as well as atom pairs (APs) has been used as explanatory variables. In this paper, ITC has been used only for predictor selection, after which ridge regression is employed to build the final predictive model. The proper leave-one-out (LOO) method of cross-validation in this scenario is to take as holdout each of the 508 compounds *before* predictor thinning and compare the predicted values with the experimental data. ITC based results obtained here are comparable to those developed earlier.

Keywords: Atom pairs, interrelated two-way clustering, mutagenicity, quantum chemical descriptors, ridge regression, topological indices.

1. INTRODUCTION

Prediction of mutagenicity is important both for drug discovery and environmental protection. If the mutagenicity of drug candidates is detected in the discovery phase of drug development, than it can lead to better and earlier decisions about the allocation of resources in the costly drug discovery process which currently needs US $400 million to 2 billion per drug [1, 2]. Regulatory agencies like the United States Environmental Protection Agency (USEPA) [3], routinely assess chemicals for their potential mutagenicity for hazard estimation. In the area of human health risk assessment of chemicals, carcinogenicity and mutagenicity are two toxicologically important end points. Because testing of a large number of chemicals in the laboratory is prohibitively expensive, alternative approaches to the bioassay were designed based on the molecular structure of chemicals as cost effective alternatives [4] for the identification of mutagens. Therefore, one recent trend in mutagenicity prediction is the use of theoretically computed descriptors in the development of models [5-7].

Different methods, like multiple regression, fuzzy logic [8], neural networks [8, 9], multistep models [10] have been used by various authors for the prediction of mutagenicity of both congeneric as well as structurally diverse sets of chemicals. All of these use easily calculated molecular descriptors, including topological indices and atom pairs (APs). One of the smaller sets of chemical mutagens studied exhaustively is the group of 95 aromatic and heteroaromatic amines originally collected by Debnath *et al.* [11]. A large number of studies has been reported in the literature on the QSARs of this set of chemicals using different classes of molecular descriptors [12-16]. In two earlier papers, Basak *et al.* [17] and Hawkins *et al.* [18] developed predictive models for mutagenicity of both a congeneric set of 95 aromatic amines and a diverse set of 508 chemicals consisting of mutagens and non-mutagens. The descriptors used included topostructural (TS), topochemical (TC), three dimensional (3-D), and quantum chemical (QC) descriptors. In our recent quantitative structure-activity relationship (QSAR) studies, we observed that the addition of calculated atom pairs (APs) to the collection of explanatory variables enhanced the quality of the models [19, 20]. So, it was of interest to investigate whether the addition of APs to the set of numerical molecular descriptors could help us in developing better models for chemical mutagenicity estimation.

We calculated a total of 2,525 descriptors for the 508 chemicals. For the selection of important independent variables from this large pool, we adapted a clustering approach first proposed by Tang *et al.* [18]. Originally applied to gene expression data obtained from microarrays, this unsupervised method, named Interrelated Two-way Clustering (ITC), iteratively selects important genes and classifies samples simultaneously. Here we substitute samples for compounds, and genes for predictors. Since in our case we have the number of predictors nearly 5 times the number of samples, it fits the $n << p$ scenario in the gene data for which it is found to perform better than K-means or Self-Organizing Maps (SOM) clustering methods [21]. After

*Address correspondence to this author at the Department of Chemistry & Biochemistry, University of Minnesota Duluth-Natural Resources Research Institute, 5013 Miller Trunk Highway, Duluth, Minnesota, 55811, USA; Tel: (218)720-4230; Fax: (218)720-4328; E-mail: sbasak@nrri.umn.edu

1875-6697/13 $58.00+.00

predictor selection through ITC, ridge regression (RR) was used to build the final QSAR model for classifying compounds as mutagen/non-mutagen. An important point to be noted here is the approach to cross-validation. If we first select important predictors and then use cross-validation to select the tuning parameter in RR, it actually uses information from the holdout compound to build the model. So it is imperative that for each holdout sample compound, we do predictor thinning using other compounds and then use RR to predict the class of this compound [22]. Following this, the effectiveness of predictor selection by ITC is compared with the results obtained by modeling and subsequent classification results on the same dataset by Hawkins *et al.* [18].

2. MATERIALS AND METHODS

2.1. The Database

The data were taken from the CRC Handbook of Identified Carcinogens and Non-carcinogens [23]. The response variable is Ames mutagenicity (which is an accurate indicator of carcinogenicity [24]), the sample available being 508 compounds classified as not mutagenic (scored 0) or mutagenic (scored 1). The set of 508 is comprised of 256 mutagens and 252 non-mutagens. Table **1** below gives an idea regarding the diversity of the chemicals in this database in terms of chemical types and functional groups.

Table 1. Major Chemical Classes (Not Mutually Exclusive) within the Mutagen/Non-Mutagen Database

Chemical Class	Number of Compounds
Aliphatic alkanes, alkenes, alkynes	124
Monocyclic compounds	260
Monocyclic carbocycles	186
Monocyclic heterocycles	74
Polycyclic compounds	192
Polycyclic carbocycles	119
Polycyclic heterocycles	73
Nitro compounds	47
Nitroso compounds	30
Alkyl halides	55
Alcohols, thiols	93
Ethers, sulfides	38
Ketones, ketenes, imines, quinones	39
Carboxylic acids, peroxy acids	34
Esters, lactones	34
Amides, imides, lactams	36
Carbamates, ureas, thioureas, guanidines	41
Amines, hydroxylamines	143
Hydrazines, hydrazides, hydrazones, traizines	55
Oxygenated sulfur and phosphorus	53
Epoxides, peroxides, aziridines	25

2.2. Calculation of Descriptors

Software packages including POLLY v.2.3 [25], Sybyl v.6.2 [26], MOPAC v 6.00 [27] and Molconn-Z [28] were used to calculate descriptors, based solely on chemical structure. The triplet indices were calculated by in-house software developed by Basak *et al.* [29] which can calculate descriptors formulated by Filip *et al.* [30]. Atom pairs are calculated by the software APProbe [31] which calculated APs following the method of Carhart *et al.* [32]. The descriptors can be classified according to their complexity and demand on computational resources. The topostructural indices make up the simplest descriptor class, encoding information related solely to the connectedness of the atoms within a molecule. The topochemical descriptors are more complex, encoding not only information related to molecular topology but also information on atom and bond types. The geometric descriptors encode three-dimensional aspects of molecular structure; and the most complex and computationally demanding quantum chemical descriptors are based on the electronic aspects of molecular structure. Table **2** gives the symbols and definition of the majority of descriptors, including geometrical and quantum chemical indices, used in this study. Atom pairs are not given in a table because of the large size. The values of all calculated TIs, 3-D descriptors, QC chemical indices, and APs are given in the supplementary material accompanying the manuscript.

2.3. Statistical Analysis

2.3.1. An Overview of Interrelated Two-Way Clustering (ITC)

The Method: This method of unsupervised analysis aims to simultaneously select important predictors as well as cluster samples into two different classes (e.g. diseased and control) in a *single iterative procedure*. For this, at first predictors are divided into different groups using some known classification method or practical considerations and then samples are classified for each predictor group independently. The idea is that if the predictors are important in class detection, the sample classifications will be identical for each predictor group. To achieve this, in each iterative step some predictors are eliminated so that sample classifications based on different predictor groups become more and more similar.

The details of the procedure are as described below:

a. *Pre-processing:* In gene expression data, the predictors (i.e. genes) having little contribution in determining the class of a sample exhibit very little change in intensity values across different samples. The pre-processing stage is used to do a preliminary filtering of such predictors. For this the predictor vectors are first normalized [33]:

$$w'_{ij} = \frac{w_{ij} - \mu_i}{\mu_i}, \text{ with } \mu_i = \frac{\Sigma_{j=1}^{m} w_{ij}}{m}$$

where w'_{ij} and w_{ij} are the normalized and crude intensity values of the i-th predictor in the j-th compound, respectively, i=1,2…n, n being the number of predictors and m the number of compounds. After this the slope of each predictor

Table 2. Symbols, Definitions and Classification of Topological Indices

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	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
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
$^h\chi$	Path connectivity index of order h = 0-10
$^h\chi_C$	Cluster connectivity index of order h = 3-6
$^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
$^h\chi_{Ch}$	Chain connectivity index of order h = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban’sJ index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
$\mathrm{DN^2S_y}$	Triplet index from distance matrix, square of graph order, and distance sum; operation y = 1-5
$\mathrm{DN^21_y}$	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$\mathrm{AS1_y}$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$\mathrm{DS1_y}$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
$\mathrm{ASN_y}$	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
$\mathrm{DSN_y}$	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
$\mathrm{DN^2N_y}$	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
$\mathrm{ANS_y}$	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$\mathrm{AN1_y}$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
$\mathrm{ANN_y}$	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
$\mathrm{ASV_y}$	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
$\mathrm{DSV_y}$	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
$\mathrm{ANV_y}$	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
$\mathrm{O_{orb}}$	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
$\mathrm{I_{ORB}}$	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
$\mathrm{IC_r}$	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
$\mathrm{SIC_r}$	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
$\mathrm{CIC_r}$	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
$^h\chi^b$	Bond path connectivity index of order h = 0-6

(Table 2) contd…..

	Topostructural (TS)
${}^{h}\chi_{C}^{b}$	Bond cluster connectivity index of order *h* = 3-6
${}^{h}\chi_{Ch}^{b}$	Bond chain connectivity index of order *h* = 3- 6
${}^{h}\chi_{PC}^{b}$	Bond path-cluster connectivity index of order *h* = 4-6
${}^{h}\chi^{v}$	Valence path connectivity index of order *h* = 0-10
${}^{h}\chi_{C}^{v}$	Valence cluster connectivity index of order *h* = 3-6
${}^{h}\chi_{Ch}^{v}$	Valence chain connectivity index of order *h* = 3-10
${}^{h}\chi_{PC}^{v}$	Valence path-cluster connectivity index of order *h* = 4-6
J^B	Balaban’s*J* index based on bond types
J^X	Balaban’s*J* index based on relative electronegativities
J^Y	Balaban’s*J* index based on relative covalent radii
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation *y* = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation *y* = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation *y* = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation *y* = 1-5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation *y* = 1-5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation *y* = 1-5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; operation *y* = 1-5
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 sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 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, $=CH_2$

(Table 2) contd…..

	Topostructural (TS)
SHavin	E-State of C atoms in the vinyl group, *=CH-*, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for *–OH*, *=NH*, $-NH_2$, *-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*, $-NH_2$, *-NH-*, *>N*, *-O-*, *-S-*, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (*y* = 2-10)
SHBinty	E-State descriptors of potential internal hydrogen bond strength (*y =2-10*)
ka_1-ka_3	Kappa alpha indices
	Electrotopological State index values for atom types: *SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, HmaxGmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH, SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb*
	Geometrical (3-D)
^{3D}W	3D Wiener number based on the hydrogen-suppressed geometric distance matrix
$^{3D}W_H$	3D Wiener number based on the hydrogen-filled geometric distance matrix
V_W	Van der Waal's volume
	Quantum Chemical (QC)
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO-1}	Energy of the second highest occupied molecular
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO+1}	Energy of the second lowest unoccupied molecular orbital
ΔHf	Heat of formation
μ	Dipole moment

vector $g_i = (w'_{i1},\ldots,w'_{im})$ with respect to the constant vector $E_{n\times 1} = (1,\ldots,1)'$ is calculated:

$$\cos(\theta) = \frac{< g_i , E >}{\| g_i \| . \| E \|} = \frac{\sum_{j=1}^{m} e_j w'_{ij}}{\sqrt{\sum_{j=1}^{m} w'^2_{ij}}\sqrt{\sum_{j=1}^{m} e_j^2}}$$

where θ is the angle between the predictor vector and the constant vector. A value close to 1 indicates the predictor intensity vector does not vary much across samples, while a value close to 0 implies near orthogonality of these two vectors, thus much deviation of the predictor expression from constant behavior across samples. In Tang *et al.* [21], a threshold of 0.9 was used to filter the normalized vectors in this step.

b. *Details of the iterative procedure:* Each iteration involves the following steps:

STEP 1- Clustering in predictor dimension: Predictors are clustered into *k* groups ($G_1,\ldots,G_k$) using a standard clustering algorithm like K-means [34, pp. 461] or SOM [34, pp. 480].

STEP 2- Clustering in sample dimension: For each predictor group G_i, samples are clustered in two groups $S_{i,a}$ and $S_{i,b}$, as per most popular experimental conditions [35].

STEP 3- Combining the two clusterings: For each predictor group two clusters are obtained, so each sample can be in any one of the two clusters for each of the *k* groups. Thus there can be 2^k possible groups of samples. For example, for *k* = 2, the 4 groups will be:

$C_1 = S_{1,a} \cap S_{2,a}$ $C_2 = S_{1,b} \cap S_{2,a}$
$C_3 = S_{1,a} \cap S_{2,b}$ $C_4 = S_{1,b} \cap S_{2,b}$

STEP 4- Obtaining heterogeneous groups: From these 2^k sample groups we select 2^{k-1} heterogeneous groups (C_s,C_t). These groups are selected in such a way that for all $\mu \in C_s$ and $\vartheta \in C_t$, if $\mu \in S_{i,p}$, $\vartheta \in S_{i,q}$ then $a \neq b$ for all *i*. In the above case for k = 2, (C_1,C_4) and (C_2,C_3) are two heterogeneous groups.

STEP 5- Sorting and reducing: For each heterogeneous group (C_s, C_t) two patterns are introduced: (0,0, …, 0,1,1, …, 1) and (1,1, …, 1,0,0, …, 0), containing $|C_s|$ (= #samples in C_s) zeros (ones) followed by $|C_t|$ ones (zeros), respectively. Vector cosines with these two patterns are calculated for predictor vectors in this heterogeneous group. The cosine values are sorted in decreasing order, and for each pattern the top one third of the predictors are kept. By merging the two sequences for the two complementary patterns we obtain a set of predictors which is reduced by at least one thirds from the predictors the iteration was started with.

Similar sets of predictors are generated for all the heterogeneous groups. Now, to select the final sequence of predictors for the iteration, leave-one-out cross validation is performed, i.e. for each heterogeneous group, select a sample, use the remaining samples to select important predictors, and use these predictors to predict the class of the withheld sample. This is repeated for all samples, and finally a cumulative error rate is obtained for each heterogeneous group. The group that has the lowest error rate has its corresponding reduced sequence of predictors selected as the set of predictors for next iteration.

c. Termination condition: We define the term Occupancy ratio as:

$$Occratio = Max \frac{|C_s + C_t|}{m}$$

where the maximum is taken over all heterogeneous groups (C_s, C_t). When the predictor clustering results based on the predictor groups are the same, one of the heterogeneous groups will contain all the samples, thus the occ-ratio will be 1. Then the reduced predictor sequence obtained in STEP 5 will be good for sample clustering. Of course, this optimal condition is hard to reach in practical situations, so a threshold value of 0.9 is used as the termination condition for iterations. To include the cases where the occ-ratio value cannot reach the threshold after many iterations, yet the remaining number of predictors becomes too small, a specific threshold of 100 predictors is used as an alternate termination condition.

Adapting the ITC in QSAR scenario: In our case, predictors take the place of genes, and samples are substituted by sample compounds. Since number of predictors is much higher than the number of compounds, this is also a case of the $n < p$ scenario in which the ITC was originally applied. The ITC is applied here in the following way:

a. We already have the classification of predictors: TS/TC/3D/QC/AP; so the clustering in gene direction was not needed.

b. For each predictor group, sample compounds were clustered using K-means ($k = 2$).

2.3.2. Transformation of Predictors

Before applying any statistical procedure, the data containing selected predictors are transformed. Because differences in magnitudes across predictors might not be of the same order, each entry x in the data is transformed as $x' = log(x+C)$ where $C = 1$ when $x > 0$, and otherwise is the negative of the largest integer less than x. For example, if $x = -1.7617$ then we take $C = 2$. After this transformation on the explanatory variables, we center and scale the response and explanatory variables by subtracting and dividing each entry by the mean and standard deviation of that column, respectively.

2.3.3. Ridge Regression

Ridge Regression (RR) [36, pp. 239] is used in place of OLS regression where significant correlation exists between different explanatory variables. For a given data (centered and scaled, no intercept term), with ***X*** being the matrix of predictors and ***Y*** the vector of responses among samples, the vector ***b*** of estimates for coefficients obtained by RR is given by

$$\mathtt{b} = (\mathtt{X'X} + k\mathtt{I})^{-1}\mathtt{X'Y}$$

where $k > 0$ is the ridge constant. A value of $k = 0$ corresponds to OLS regression, while as k grows to infinity, the RR estimates shrink towards 0.

Methods of choosing k suitably include using leave-one-out prediction sum of squares (PRESS) statistic and Generalized Cross-Validation (GCV) [36, pp. 253]. In each of these the fitted cross-validation score is calculated for a range of values, and we choose as k the value for which this score is minimum. Finally the predictive ability of the model is assessed by cross-validated classification score, which is obtained the following way:

a. Remove one compound from the data set. Fit a RR model with the rest of the compounds by choosing an optimal k and obtaining the corresponding vector of coefficients.

b. Obtain the predicted response for the holdout compound by multiplying this vector of coefficients with the predictor vector of this compound. If the value obtained is greater than a previously fixed cutoff value, say, c, then take the predicted mutagenicity score as 1, otherwise take it 0.

c. Continue for all the sample compounds. Now by comparing the true and predicted mutagenicity classifications we obtain the cross-validated classification score.

2.3.4. Naïve and Proper Cross Validation

When the setup includes a step to select important variables before performing the model-building, it is imperative that the cross-validation is performed at the proper stage. It must be ensured that no information from the holdout compound is used in any way for predicting its response through cross-validation. That is why first doing predictor selection and then using cross-validation to obtain the model is not the proper way to do cross-validation [22], because then the first step of thinning involves the holdout compound as well. So, in our situation we do the

Table 3. Results of RR Analyses of 508 Mutagens and Non-Mutagens Using Calculated Descriptors

Type of Predictors in Model	Model Description	No. of Predictors	Type of Cross Validation	Correct Classification %	Sensitivity	Specificity
TS+TC	Ridge regression without descriptor thinning	298	Leave-one-out CV	76.97	83.98	69.84
TS+TC+3D+QC	Ridge regression without descriptor thinning	307	Leave-one-out CV, done by Hawkins *et al.* [18]	77.17	84.38	69.84
TS+TC+AP	RR with ITC thinning (after first iteration)	203	Two-deep CV	78.35	84.38	72.22

cross-validation by omitting each compound, separately perform the ITC clustering routine and choose the ridge constant *k* for all of them, and then predict mutagenicity scores of the compounds.

3. RESULTS AND DISCUSSION

The objectives of our work were two-fold:

a) To investigate how far the addition of AP descriptors to the set of TS+TC+3D+QC indices increases the quality of models for predicting mutagenicity of chemicals, and

b) To adapt the ITC method, originally developed for application in gene expression data analysis, in the selection of a subset of explanatory variables from a large pool of descriptors.

All the analyses were performed in MATLAB, version R2008a [37]. We followed a hierarchical approach by first performing RR on only TS and TC descriptors. RR results involving TS, TC, 3D and QC descriptors on the same set of chemicals were obtained by Hawkins *et al.* [18] previously. As we can see in the results in Table **3**, the predictive ability before and after including the 3D and QC descriptors is almost the same, with the TS+TC model having a loss of sensitivity (i.e. correct prediction percentage for mutagens) of 0.4 and the same specificity (i.e. correct prediction percentage for non-mutagens), translating to misclassification of only one mutagen compared to the TS+TC+3D+QC model. This is in line with earlier hierarchical QSAR (HiQSAR) studies of Basak *et al.* [12, 14, 15, 17, 38, 39] for various sets of physicochemical property, bioactivity, and toxicity data that 3-D and QC descriptors make very little or no improvement in model quality after the use of TS and TC descriptors.

Taking this into account, and also the fact that including 5 types of descriptors in the ITC algorithm would result in 2^4 = 16 heterogeneous groups and thus significant increase of computational load, we decided to perform the ITC thinning on TS+TC+AP descriptors instead of the TS+TC+3D+QC+AP descriptors. In the first iteration, the occ-ratio reached 0.89 and the descriptors gave better predictive scores than by RR on the full TS+TC+3D+QC set. Going into the second iteration, the occ-ratio improved slightly to 0.9075, terminating the algorithm, but the number of predictors was diminished, resulting in significant decrease of predictive ability. Because of this the model built from the set of descriptors obtained after the first iteration was taken as final.

Cutoff for predicted mutagenicity was taken as c = 0.5 for all methods. The results are summarized in the table above

The set of predictors obtained after ITC thinning contained 57 topostructural, 101 topochemical and 45 atom-pair descriptors. Although this model contained fewer predictors, it gave a 2% increase of specificity than RR using first 4 types of descriptors. It is also to be noted that ITC analysis reported here did variable selection from a large pool of descriptors contained in the TS+TC+AP set of explanatory variables.

12 descriptors (6 topochemical, 6 atom-pair) were found to have $|t|$-ratios that are significant at 95% confidence level. Interestingly, no topostructural descriptors were among these 12. One possible interpretation could be that of the compound set containing a large variety of compounds of different structural classes and the structural information encoded by just the connectivity of atoms without any consideration of the atomic characters or bonding patterns were not enough to predict mutagenicity efficiently. Therefore all the 6 influential TIs were electrotopological state indices. The $|t|$-ratios and names and types of these descriptors are as in Table **4**:

Table 4. Descriptors with Significant $|t|$-Values from the RR Model Obtained Using Descriptors Selected by ITC

Descriptor Name	$|t|$-Ratio	Descriptor Class
SsCH3	3.7555	TC
$O.X_1$-3-$O.X_1$	3.4157	AP
NX_3-3-$S.X_1$	2.8252	AP
$C.X_3$-2-NX_3	2.8246	AP
SaaNH	2.7164	TC
SdS	2.4749	TC
CX_1-3-OX_2	2.4209	AP
OX_2-2-$S..X_4$	2.2023	AP
StsC	2.1987	TC
$C.X_3$-2-$C.X_3$	2.1623	AP
SsNH2	2.1488	TC
SDsssP	2.0329	TC

4. CONCLUSION

The main objective of this paper was to study the effectiveness of ITC method vis-à-vis RR technique in the prediction of mutagenicity/ non-mutagenicity of a diverse set of chemicals. Results show that ITC can do effective variable selection from a large pool of calculated descriptors. Predictive models developed from the ITC-derived descriptors compare reasonably well with those developed using the RR method. Further studies with other sets of bioactive chemicals, as well as employing cross-validation with a training set of 10-20% of samples to protect against overfitting and comparing with the results obtained by leave-one out CV are needed to characterize the relative effectiveness of these methods in QSAR development.

CONFLICT OF INTEREST

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

ACKNOWLEDGEMENTS

This paper is dedicated to Professor Lemont B. Kier on his 80th birthday. The authors would like to thank Professor Douglas Hawkins, School of Statistics, University of Minnesota, for his help in data analysis carried out in this paper. One of the authors (SBM) would like to thank Satadip Saha for helping in running the codes.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's web site along with the published article.

REFERENCES

[1] Adams, C.; Brantner, V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? *Health Aff. (Millwood)*, **2006**, *25*, 420-428.

[2] DiMasi, J.; Hansen, R.; Grabowski, H. The price of innovation: new estimates of drug development costs. *J. Health Eco*., **2003**, *22*, 151-185.

[3] United States Environmental Protection Agency; http://www.epa.gov/nrmrl/std/cppb/qsar/index.html

[4] Benigni, R.; Bossa, C.; Tcheremenskaia O.; Giuliani A. Alternatives to the carcinogenicity bioassay: *in silico* methods, and the *in vitro* and *in vivo* mutagenicity assays. *Expert Opin. Drug Metab. Toxicol.*, **2010**, *6*, 1-11.

[5] Bacha, P.A.; Gruver, H.S.; Hartog, B.K.D.; Tamura S.Y.; Nutt, R.F. Rule extraction from a mutagenicity data set using adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.,* **2002**, *42,* 1104-1111.

[6] Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Müller, K. Benchmark data set for *in silico* prediction of Ames mutagenicity. *J. Chem. Inf. Model,* **2009**, *49,* 2077-2081.

[7] Helma, C.; Cramer, T.; Kramer, S.; DeRaedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J. Chem. Inf. Comput. Sci.,* **2004**, *44,* 1402-1411.

[8] Benfenati, E. The CAESAR project for *in silico* models for the REACH legislation. *Chem. Cent. J.*, **2010**, *4*, 11.

[9] Fjodorova, N.; Novic, M. Some findings relevant to the mechanistic interpretation in the case of predictive models for carcinogenicity based on the counter propagation artificial neural network. *J. Comput. Aided Mol. Des.*, **2011**, *25*, 1159-1169.

[10] Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem. Cent. J.*, **2010**, *4*, S2.

[11] Debnath, A.K.; Debnath, G.; Shusterman, A.J.; Hansch, C. A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in *Salmonella* typhimurium TA98 and TA100. *Environ. Mol. Mutagen*, **1992**, *19*, 37-52.

[12] Basak, S.C.; Mills, D. Quantitative molecular similarity analysis (QMSA) methods for property estimation: A comparison of property-based, arbitrary, and tailored similarity spaces. *SAR QSAR Environ. Res.,* **2001**, *12*, 481-496.

[13] Cash, G.G. Prediction of the genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mutat. Res.,* **2001**, *491*, 31-37.

[14] Basak, S.C.; Gute, B.D.; Grunwald, G.D. In: *Quantitative Structure-activity Relationships in Environmental Sciences VII*, F. Chen; G. Schüürmann, Eds.; SETAC Press: Pensacola, FL, **1998**; pp. 245-261.

[15] Basak, S.C.; Mills, D.; Balaban, A.T.; Gute, B.D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 671-678.

[16] Maran, U.; Karelson, M.; Katritzky, A.R. A Comprehensive QSARs treatment of the genotoxicity of heteroaromatic and aromatic amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3-10.

[17] Basak, S.C.; Mills, D.; Gute, B.D.; Hawkins, D.M. In: *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, R. Benigni, Ed.; CRC Press: Boca Raton, FL, **2003**; pp. 207-234.

[18] 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.

[19] Basak, S.C.; Zhu Q.; Mills, D. Prediction of anticancer activity of 2-phenylindoles: comparative molecular field analysis versus ridge regression using mathematical molecular descriptors. *Acta Chim. Slov.,* **2010**, *57*, 541-550.

[20] Basak, S.C.; Zhu, Q.; Mills, D. Quantitative structure-activity relationships for anticancer activity of 2-phenylindoles using mathematical molecular descriptors. *Curr. Comput. Aided Drug Des.*, **2011**, *7*, 98-108.

[21] 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.

[22] Hawkins, D.M.; Kraker, J.J.; Basak, S.C.; Mills, D. *QSPR checking and validation: a case study with hydroxy radical reaction rate constant. SAR QSAR Environ. Res.*, **2008**, *19*, 525-539.

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

[24] McCann, J.; Choi, E.; Yamasaki, E.; Ames B.N. Detection of carcinogens as mutagens in the Salmonella/microsome test: assay of 300 chemicals. *Proc. Natl. Acad. Sci. USA*, **1975**, *72*, 5135-5139.

[25] Basak, S.C.; Harriss, D.K.; Magnuson, V.R. *POLLY v. 2.3,* Copyright of the University of Minnesota, **1988**.

[26] Tripos Associates, Inc. *Sybyl Version 6.2,* St. Louis, MO, **1995**.

[27] Stewart, J.J.P. *MOPAC Version 6.00, QCPE #455*, Frank J Seiler Research Laboratory: US Air Force Academy, CO, **1990**.

[28] Hall Associates Consulting, *Molconn-Z Version 3.50*, Quincy, MA, **2000**.

[29] Basak, S.C.; Grunwald, G.D.; Balaban, A.T. *TRIPLET,* Copyright of the Regents of the University of Minnesota, **1993**.

[30] Filip, P.A.; Balaban, T.S.; Balaban, A.T. A New Approach for Devising Local Graph Invariants: Derived Topological Indices with Low Degeneracy and Good Correlation Ability. *J. Math. Chem.,* **1987**, *1*, 61-83.

[31] Basak, S.C.; Grunwald, G.D. *APProbe*, Copyright of the University of Minnesota, **1993**.

[32] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies. *J. Chem. Inf. Comput. Sci.,* **1985**, *25*, 64-73.

[33] Schuchhardt, J.; Beule, D.; Malik, A.; Wolski, E.; Eickhoff, H.; Lehrach, H.; Herzel, H. Normalization strategies for cDNA microarrays. *Nucl. Acids Res.*, **2000**, *28*, e47.

[34] Hastie T.; Tibshirani R.; Friedman J. *The Elements of Statistical Learning - Data Mining, Inference and Prediction*, 2nd ed.; Springer: New York, NY, **2008**.

[35] Brazma, A.; Vilo, J. Gene expression data analysis. *FEBS Lett.*, **2000**, *480*, 17-24.

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

[37] Mathworks Inc. *MATLAB Version 7.6 (R2008a)*, **2008**.

[38] Gute B.D.; Basak, S.C. Predicting acute toxicity (LC_{50}) of benzene derivatives using theoretical molecular descriptors: A hierarchical approach. *SAR QSAR Environ. Res.*, **1997**, *7*, 117-131.

[39] Basak, S.C.; Gute, B.D.; Grunwald, G.D. Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 651-655.

Received: August 3, 2012 **Revised: September 13, 2013** **Accepted: October 4, 2013**