Ridge Regression Fitting and Diagnostics in

Quantitative Structure--Activity Relationship Modeling

Douglas M. Hawkins [a], Subhash C. Basak [b], Denise Mills [b]

[a] School of Statistics, 313 Ford Hall, 224 Church Street S. E., University of Minnesota

Minneapolis, Minnesota 55455, USA

[b] Natural Resources Research Institute, University of Minnesota Duluth

5013 Miller Trunk Highway, Duluth, MN 55811, USA

Corresponding Author:
Douglas M. Hawkins
School of Statistics
313 Ford Hall, 224 Church Street S. E.
University of Minnesota
Minneapolis, Minnesota 55455

Phone: (612) 624-4166
Fax: (612) 624-8868
Email: doug@stat.umn.edu

Abstract

Quantitative Structure-Activity Relationship (QSAR) modeling is a vital tool in leveraging limited experimental measurements of chemicals' capacity for environmental damage bye allowing predictions to be made of ~~the~~ such properties as toxicity or mutagenicity of compounds for which experimental data are unavailable. All common QSAR modeling methods –regression, neural nets, *k*-nearest neighbors – are 'linear smoothers'. Their predictions of a compound's activity are weighted averages of the activities in the calibration data with weights dependent on the descriptors. While the methods have been studied extensively, a vital but overlooked area is 'case diagnostics', pointers to compounds that are poorly fitted, or are unusually influential in fitting the model. This is particularly true where the measured activity is binary – present or absent. We discuss useful numerical and graphical diagnostics, and illustrate their use, particularly that of the FF plot, in the context of a ridge regression of a data set with 508 compounds and 307 structural descriptors used to predict Ames mutagenicity.

Keywords: Linear model; Diagnostics; Leverage; Influence; Residuals; Mutagenicity

1. Introduction

An important problem in environmental toxicology and risk assessment of chemicals is the prediction of adverse effects of pollutants on human and environmental health from their structure (Auer et al., 1990). There are more than 82,000 substances in current use as chemicals in commerce in the United States and the list, the Toxic Substances Control Act (TSCA) Inventory, is growing by 2,000 to 3,000 per year. Most of the TSCA chemicals have no data necessary for their hazard estimation. For example, more than 50% have no available chemical, physical or biological test data at all, and only 15% have mutagenicity, genotoxicity or chronic toxicity data. So, the usual practice of property-property correlation where a suite of simple or easily available properties is used to predict more a complex toxicity endpoint (Hansch et al., 1995) is not useful in the prediction of potential hazards of substances in the TSCA Inventory ~~and also~~ or the new substances being submitted to the United States Environmental Protection Agency every day. The same is true of the myriad of chemicals detected in the more than 1,000 Superfund sites around USA (Auer, *et al.*, 1990). A viable solution to this quagmire is the prediction of toxicity and toxicologically relevant properties of chemicals in commerce from their molecular structure without the input of any laboratory data. To this end, various quantitative structure-activity relationship (QSAR) models using calculated molecular descriptors have been reported in the literature (Basak et al., 2002a;

Basak et al., 2000b; Basak et al., 2002b; Eldred and Jurs, 1999; Eldred et al., 1999; Katritzky and Tatham, 2001; Serra et al., 2001).

Mutagenicity is an important toxicological endpoint both for environmental chemicals and potential therapeutic agents (Benigni and Giuliani, 1996; Benigni et al., 2000; Blake et al., 1990; Debnath et al., 1992; Frierson et al., 1986; Moyer and Jurs, 1990). Carcinogens fall into three major categories: genotoxic, epigenetic, and foreign body carcinogens. The first category of carcinogens either themselves or after metabolic activation alkylate DNA and are mutagenic in nature. Hence mutagenicity of chemicals is an important factor in assessing the hazardous effects of chemicals on human and environmental health. Therefore, it is not surprising that various authors have attempted to predict mutagenicity of chemicals from their structure as well as available experimental data (Basak et al., 1986; Basak and Grunwald, 1995; Basak et al., 1998; Basak et al., 1999; Basak and Mills, 2001; Basak et al., 2002c; Basak et al., 2001; Benigni and Giuliani, 1996; Benigni, *et al.*, 2000; Blake, *et al.*, 1990; Debnath, *et al.*, 1992; Frierson, *et al.*, 1986; Moyer and Jurs, 1990).

1.2 QSAR Model Development

Biological activity may be measured on a variety of scales. The best situation is where the activity is a continuous value – a percent suppression for example. Continuous dependent variables allow for fine gradations of activity and make it possible to fit QSAR models with quite modest amounts of information. The least discriminating measure is a binary – presence or absence. While good models can be fitted to binary activity measures, it tends to require a larger calibration set than suffices with a continuous activity measurement.

The structural information used as predictors in QSAR can also be of different types. Angles and bond energies are examples of continuous predictors. Predictors may be binary – for example presence or absence of a benzene ring. Some predictors – for example the number of nitrogen atoms – are integer-valued, and are intermediate in information content between binary and continuous predictors.

QSAR modeling often involves large numbers of predictors: using several hundred is quite common. Various approaches followed for the modeling include:

1. Linear modeling, in which the prediction is of the form $\Sigma_j b_j x_j$ where the x_j are the predictors and the b_j are fitted coefficients. This includes conventional multiple regression (where applicable), partial least squares (PLS), principal component regression (PCR) and ridge regression (RR).
2. Nonlinear modeling methods such as neural nets, *k*-nearest neighbor and nonlinear regression. We will not say more about these methods here.
3. Subset methods that seek to find some small subset of the structural variables that captures all their predictive information
4. For a binary activity measure, discriminant analysis and logistic regression.
5. If the number of compounds available for calibration is sufficiently large, recursive partitioning is attractive (Hawkins et al., 1997).

Subset methods are intuitively attractive because of the appeal of Occam's razor. However they are inherently biased, tend to vastly overestimate their own accuracy, and have been found to give worse predictions than full-sample methods. Better predictions appear to be found (Frank and Friedman, 1993) by~~t~~ methods that retain all features but take some steps to address overdetermination problems.

The linear modeling approaches at first sight seem hopelessly naïve in their assumption of a global linear relationship; however they have been found to give sturdy, quite accurate models provided the true relationship with each predictor is monotonic. They can even, under some circumstances, give an unbiased picture of nonlinear relationships (Li and Duan, 1989).

Linear discriminant analysis (LDA) for a binary dependent turns out to be exactly equivalent to coding the dependent as 0/1 and using linear regression. Quadratic discriminant analysis and logistic regression are not particularly attractive in the context of very large numbers of predictors – QDA because of the huge number of parameters to estimate, and logistic regression because of the problem of exact separating hyperplanes.

The method we have used in the following discussion is ridge regression (RR) using all the descriptors that have been recorded. RR is attractive among the linear methods in that it has some established good theoretical properties, though it is substantially more computationally intensive than is say PLS.

2. Fitting and verifying a QSAR model

Write *n* for the number of compounds available for calibration, and *p* for the number of predictors. All the commonly used methods involve the determination of constants in the QSAR model. For example, in any of the linear modeling approaches, suitable values for the coefficients of the predictors need to be found. Fitting a RR proceeds as follows:-

1. Standardize ("autoscale") each predictor by subtracting the predictor's mean and dividing by its standard deviation. Prior to even this step, a decision may be made to convert the dependent and/or predictor variables to some other measurement scale. For example, a common heuristic observation is that if any necessarily positive variable varies by at least two orders of magnitude across the data set, then it should probably be transformed to a log scale. Any such preliminary transformation to a log or some other scale will be taken as a given.
2. Then write **X** for the $n \times p$ 'design' matrix of the predictors with each row being a compound and each column a descriptor. Write **Y** for the vector of activities (possibly following a log or other transformation)
3. The RR coefficient vector **b** is given by

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

where *k* is a non-negative constant known as the 'ridge' constant.

The value *k*=0 corresponds to conventional least squares regression. Fitting a RR involves choosing a suitable value of *k*, after which the coefficients follow at once. There are two standard ways of finding *k* – leave-one-out cross-validation (CV), and generalized cross-validation (GCV). Each involves searching over possible *k* values to pick that value giving the best fit. In a CV, you omit each compound in turn, use the remaining compounds and the test value of *k* to find the coefficient vector **b**, then use that value of the coefficient vector to predict the held-out case. The difference between the case's *y* and its hold-out prediction is called the case's *predicted residual*. The value of *k* chosen is that value minimizing the sum of squares of the predicted residuals.

Note that CV leads to fitting a total of *n* ridge regressions for each *k* –one for each of the *n* subsamples obtained by the temporary removal of a case – along with the final RR using all cases for ~~the selected value of~~each *k*.

GCV (Golub et al., 1979), uses a mathematical argument to simulate the prediction of future unknowns and thereby avoids the *n* temporary removals and refittings by using a calculation involving just the full-sample RR. This leads to a substantially faster calculation and also to a theoretically better choice of *k*, but at the cost of losing the transparency that CV has in that the predicted residuals visibly escape any bias of tuning of the model to the case being predicted.

The second part of the fitting involves model assessment – estimating how well the model will perform in predicting future compounds to which it is applied. Cross-validation is a clearly attractive method of doing this assessment in that the predicted residual found in cross validation simulates the use of the fitted model in making a future prediction. However for CV to provide a reliable picture of the model's validity in future predictions, it is vital that the prediction of the hold-out sample not involve the results of the holdout sample *in any way*.

Suppose for example, RR is used, and CV is used to select *k*. Then a proper CV to assess the model fit would proceed as follows:-

1. Temporarily remove compound *i* from the data set, leaving *n*-1 compounds. To fit a RR to these *n*-1 compounds we require the value of *k* for those compounds. We would get this by looking at various choices of *k*. For each test *k*, successively remove each of the *n*-1 compounds, fit the RR to the remaining *n*-2 to predict this second-level holdout, and in this way find the appropriate *k* for these *n*-1 compounds. Fit the overall RR using this *k* and use it to predict the first-level holdout compound.
2. Doing this for each of the *n* compounds we get truly honest holdout predictions and predicted residuals.
3. Assess the quality of the fit from these predicted residuals.
4. Go back and pick the *k* for the full sample.

This involves a large number of regression calculations – *n*(*n*-1) for each *k*. Using GCV is preferable in that it avoids the inner loop of *n*-1 regressions, to pick the hold-out *k* values, though the outer cross-validation step still requires *n* leave-one-out fits in addition to the ‘global’ fit.

For other fitting methods such as PLS or the nonlinear fitting methods, the details of the inner fitting differ, but the need for a proper outer cross-validation remains.

3. Case diagnostics

In any model fitting work, there is concern about ‘cases’ – in QSAR work, compounds – that call for special attention. Three properties of a case matter:-

1. A case’s leverage measures how atypical its descriptors are. If for example one of the descriptors used is the number of sulfur atoms in the molecule and only one of the compounds in the calibration set contains sulfur, then this compound will have enormous leverage. It and it alone provides all information about the relevance of sulfur in predicting activity. If there are only two compounds that contain any sulfur the situation is less extreme, but both compounds will have very high leverage. High leverage cases are important because much is riding on their observed activity.
2. More important than~~t~~ leverage though is influence. A case is influential if its removal from the calibration data set leads to a substantial change in the fitted coefficient vector. Influence is commonly measured by “Cook’s distance”, which measures the overall change in the regression coefficients when the case is removed. Note that a case can be high leverage but not influential; this

happens~~will be the case~~ if the model fitted to the remaining cases predicts the case accurately.

3. A case is outlying if its measured activity is very different from that predicted by the other cases.

These diagnostics and their computation in a RR context are discussed in Hawkins and Yin (2002).

4. QSAR of diverse mutagens

The data were taken from the CRC Handbook of Identified Carcinogens and Noncarcinogens (Soderman, 1982). The dependent variable is Ames mutagenicity, 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 nonmutagens. Table 1 illustrates the diversity of the chemicals in this database. Software packages including POLLY v.2.3 (Basak et al., 1988), Triplet (Basak et al., 2000a; Filip et al., 1987), Sybyl v.6.2 (Tripos Associates, Inc., 1995), MOPAC v 6.00 (Stewart, 1990), and Molconn-Z (2000) were used to calculate the molecular descriptors, based solely on chemical structure, which can be classified according to complexity and demand for 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 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. Included in the set of descriptors calculated for use in the current study are the connectivity indices (Kier et al., 1976; Randic, 1975), path length descriptors (Kier and Hall, 1986), information theoretic (Basak, 1999; Raychaudhury et al., 1984) and neighborhood complexity indices (Basak, 1999), electrotopological state indices (Hall et al., 1991; Kier and Hall, 1999b; Kier et al., 1991), kappa shape indices (Kier and Hall, 1999a), and the Triplet indices (Filip et al., 1987). The full set of predictors is that listed and discussed in Basak and Mills (2001). ~~*et al* (xxxx), with three exceptions:~~

~~(1) Two descriptors used in that study were not used in the current study, namely, number of chlorine atoms (NoCl) and the hydrogen bonding parameter, HB_1.~~

~~(2) Three additional geometrical parameters were used in the current study, namely, van der Waals volume (V_W), the 3D Wiener number based on the hydrogen-suppressed geometric distance matrix (${}^{3D}W$), and the 3D Wiener number based on the hydrogen-filled geometric distance matrix (${}^{3D}W_H$).~~

~~(3) Six semi-empirical quantum chemical descriptors were used in the current study including energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital (E_{HOMO-1}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second lowest unoccupied molecular orbital (E_{LUMO+1}), heat of formation (ΔHf), and dipole moment (μ).~~

From the full set of descriptors, those with missing values for any of the compounds were eliminated, as were those that had constant values for all compounds in the data set.

In addition, one of every pair of perfectly correlated descriptors was removed. This left a total of 307 molecular descriptors, all of which were used in the ridge regression.

The dependent variable takes on only the values 0 and 1, but the mutagenicity score – the predictions from the model – can in principle take on any real values, positive or negative. Classification of future untested compounds involves fixing some cutpoint *c*. Any compound whose mutagenicity score exceeds *c* will be flagged as a suspected mutagen; those whose score is below *c* will be flagged as not suspicious. It is not obvious how to fix the cutoff *c*. A helpful diagnostic for this is the 'receiver operating characteristic' or ROC curve (Figure 1). To create this curve, vary *c* continuously over all real values. For each *c* count

TP (true positive) – the proportion of the mutagens whose score is above *c*; and

FP (false positive) – the proportion of non-mutagens whose score is above *c*.

The ROC is a plot of FP (horizontal) versus TP (vertical). The TP is also known as the 'sensitivity' and 1-FP is known as the 'specificity'.

The ROC curve makes explicit the tradeoff between true and false positives. All ROC curves go from the point (0,0) to the point (1,1). An ideal ROC curve rises very steeply on the left and then goes close to horizontally. The ROC curve for this data set, shown as Figure 1, is close to horizontal on the right and fairly steep on the left. For example the point (0.15, 0.63) shows a choice of *c* that will detect 63% of the mutagens while only falsely accusing 15% of the non-mutagens. Going to a 20% false positive rate increases the sensitivity to 70% of the mutagens, a 1.4 to 1 rate of recovery. Identifying 90% of the mutagens requires a false positive rate of 40%. The last and very flat portion of the curve shows that identifying the last 10% of the mutagens comes at high cost in false positives.

The default cutoff is *c*=0.5. This default gives the 'confusion matrix'

		True status		
		Mutagenic	Non-mutagenic	Total
Classification	Suspect	216	76	292
	Not suspect	40	176	216
	Total	256	252	508

giving sensitivity 84% and specificity 70%.

5. Mutagenicity case diagnostics

This default sensitivity of 84% with specificity 70% evidences a good ability to screen for mutagens in a diverse compound library. But this figure actually understates the model's potential. The library very likely includes some 'problem' compounds – for example compounds with disproportionate influence; and 'outliers' whose reported activity is implausibly different from that of structurally similar compounds. Closer study of such compounds followed by refinement in the model and perhaps changes in the library may lead to further performance improvements. Identifying such atypical compounds however presents a huge challenge.

In a conventional regression with a continuous dependent variable, an invaluable plot is that of residuals versus fitted values (Cook and Weisberg, 1999). Where the dependent is binary though this plot is much less informative – it comprises two parallel stripes of points. Replacing it is a close cousin – the FF (fitted-fitted) plot (Collett, 1991;

Olive, 1998). For this, we compute for each compound the predicted *y* made using the full data set, and that made in the cross-validation step in which the compound is omitted and the prediction made using the remaining *n*-1 compounds and plot one prediction against the other.~~.~~

Figure 2 shows the FF plot of the example data set. The mutagens and non-mutagens are shown with different plot symbols. The plot shows several interesting features.

- Nearly all the points lie very close to the 45 degree identity line. This means that the full-sample and the leave-one-out predictions are for the most part very close.
- At a finer degree of resolution, the mutagens and non-mutagens seem to cluster around two parallel but slightly separated lines. This is because removing a mutagen in the cross-validation changes the intercept by a different amount than removing a non-mutagen.
- A minority of points are far from the identity line. The prediction of such a point changes substantially when you remove it from the data set and predict it using the remaining cases. If an outlier is defined as a case whose value is much different than what you would expect based on the other cases, then these cases fit the definition of outliers, and they should be pulled out of the data set for closer investigation.

We can identify these anomalous points by calculating the difference between their full-sample and jack-knifed fitted values. It is also informative to plot this difference against the Cook’s distance used to measure the influence of each case (Figure 3). This plot shows in yet another way that there is a handful of ‘problem’ compounds that are not well predicted by the other compounds, and whose presence or absence in the calibration data has a disproportionate effect on the fitted model. Table 2 lists the 21 compounds out of the total of 508 with the largest influence, along with their Cook’s distance. When one treats a congeneric set of chemicals in a QSAR model, and some of them are outliers or have large influence on the model, an explanation is often sought from the structural side of the molecules. But the set of mutagens treated in this paper does not belong to any particular structural class. Moreover, they are represented by a large number of diverse molecular descriptors. Therefore, a structure-based explanation of the anomalous behavior of influential chemicals is not apparent. It should also be noted that outliers may be the result of experimental error, inter-laboratory variability, or chemical impurities. This further complicates matters, however, it is ~~hoped~~ anticipated that continued study will provide some insight.

It is of some interest to know the effect of removing these anomalous cases on the fit to the remaining cases. As an indication of this, repeating all the calculations on the remaining 487 cases and applying the default *c*=0.5 to classify suspect from non-suspect compounds gives the confusion matrix:

		True status		
		Mutagenic	Non-mutagenic	Total
Classification	Suspect	212	59	271
	Not suspect	34	182	216
	Total	246	241	487

giving sensitivity 86% and specificity 76%, a worthwhile ~~n~~ improvement on the diagnostic performance fitting the model to ~~using~~ all 508 compounds.

6. Discussion

The QSAR model using molecular descriptors predicts mutagenicity with high sensitivity and specificity – 84% and 70% using the default cutoff. This probably understates the model's actual capability since the data set almost inevitably has a number of 'problem' compounds such as compounds whose recorded mutagenicity is incompatible with that of others with very similar structure; compounds of a chemical class so different that a common model cannot accommodate both classes. In a large high-dimensional data set such as this, identifying these problem compounds is a daunting challenge.

The Pareto principle suggests that a high proportion of problems comes from a small proportion of cases. In the context of QSAR modeling, it is not uncommon to find that the large majority of compounds are unexceptional; that they have near neighbors and conform to what these neighbors would lead you to expect. Attention can then be focused on the small proportion of compounds for which this is not the case, since these compounds are disproportionately informative. High leverage compounds are those that do not live in well-populated structural neighborhoods, and they therefore have a disproportionately large effect on what the QSAR model has to say about future compounds in their neighborhoods. Since much is riding on the accuracy of the data in these compounds, they should be checked carefully.

Influential cases are compounds whose inclusion or exclusion from the calculations made in fitting the QSAR has a disproportionate effect on the fitted model. They are a cause for immediate concern since they are evidence of an unhealthy level of dependence of the overall analysis on these few cases. The data for influential cases should be checked for accuracy and, if possible, other compounds with similar structure should be tested and added to the data base to resolve the question of which model is the more accurate – that including the influential case or that excluding it.

Outliers – compounds whose activity differs markedly from what the other compounds would lead one to expect – may or may not be influential. They too should be checked since data capture errors are a common source of severe outliers. Regression diagnostics thus provide an easy, automated quality screen for QSAR data sets. If outliers~~they~~ are not the result of errors, then they are indications of an imperfect model and so may be important 'canaries in the coal mine.'

In the mutagenicity data, out of 508 compounds, 487 were conformist and need no further close attention. The remaining 21 were more individualistic and need closer study. Removing them leads to more accurate predictions in the remaining compounds. Also potentially valuable (though this is a possibility we do not explore here) a more detailed study may pinpoint reasons why the influential compounds'~~ir~~ activity differs so markedly from the model predictions. If this is due to some experimental error, the data set can be enhanced; if it is not, then potentially much can be learned about directions for improvement in the model.

Acknowledgement

This is contribution number ~~XXX~~346 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grant F49620-01-0098 from the United States Air F~~f~~orce and Cooperative Agreement Number 572112 from the Agency for Toxic Substances and Disease Registry.

References

Auer, C. M., Nabholz, J. V., Baetcke, K. P., 1990. Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure-activity relationships (SAR) under TSCA, Section 5. Environ. Health Perspect. 87 183-197.

Basak, S. C., Balaban, A. T., Grunwald, G. D., Gute, B. D., 2000a. Topological indices: Their nature and mutual relatedness. J. Chem. Inf. Comput. Sci. 40, 891-898.

Basak, S. C., Balasubramanian, K., Gute, B. D., Mills, D., Gorczynska, A., Roszak, S., 2002a. Prediction of cellular toxicity of halocarbons from computed chemodescriptors: A hierarchical QSAR approach. J. Chem. Inf. Comput. Sci., submitted.

Basak, S. C., 1999. Information theoretic indices of neighborhood complexity and their applications. In: Devillers, J., Balaban, A. T. (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR Gordon and Breach Science Publishers, The Netherlands, pp. 563-593.

Basak, S. C., Frane, C. M., Rosen, M. E., Magnuson, V. R., 1986. Molecular topology and mutagenicity: A QSAR study of nitrosamines. IRCS Med. Sci. 14, 848-849.

Basak, S. C., Grunwald, G. D., 1995. Predicting mutagenicity of chemicals using topological and quantum chemical parameters: A similarity based study. Chemosphere 31, 2529-2546.

Basak, S. C., Grunwald, G. D., Gute, B. D., Balasubramanian, K., Opitz, D., 2000b. Use of statistical and neural net approaches in predicting toxicity of chemicals. J. Chem. Inf. Comput. Sci. 40, 885-890.

Basak, S. C., Gute, B. D., Grunwald, G. D., 1998. Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals. In: Chen, F., Schuurmann, G. (Eds.), Quantitative Structure-activity Relationships in Environmental Sciences VII SETAC Press, Pensacola, FL, pp. 245-261.

Basak, S. C., Gute, B. D., Grunwald, G. D., 1999. Assessment of the mutagenicity of aromatic amines from theoretical structural parameters: A hierarchical approach. SAR QSAR Environ. Res. 10, 117-129.

Basak, S. C., Harriss, D. K., Magnuson, V. R., 1988. POLLY, Version 2.3, Copyright of the University of Minnesota.

Basak, S. C., Mills, D., 2001. Prediction of mutagenicity utilizing a hierarchical QSAR approach. SAR QSAR Environ. Res. 12, 481-496.

Basak, S. C., Mills, D., Gute, B. D., Grunwald, G. D., Balaban, A. T., 2002b. Applications of topological indices in property/bioactivity/toxicity prediction of chemicals. In: Rouvray, D. H., King, R. B. (Eds.), Topology in Chemistry: Discrete Mathematics of Molecules Horwood Publishing Limited, Chichester, England, pp. 113-184.

Basak, S. C., Mills, D., Gute, B. D., Hawkins, D. M., 2002c. Predicting mutagenicity of congeneric and diverse sets of chemicals using computed molecular descriptors: A hierarchical approach. In: Benigni, R. (Ed.) Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens CRC Press, Boca Raton, FL, pp. in press.

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

Benigni, R., Giuliani, A., 1996. Quantitative structure-activity relationships (QSAR) of mutagens and carcinogens. Med. Res. Rev. 16, 267-284.

Benigni, R., Giuliani, A., Franke, R., Gruska, A., 2000. Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. Chem. Rev. 100, 3697-3714.

Blake, B. W., Enslein, K., Gombar, V. K., Borgstedt, H. H., 1990. Salmonella mutagenicity and rodent carcinogenicity: Quantitative structure-activity relationships. Mutation Res. 241, 261-271.

Collett, D., 1991. Modelling Binary Data. Chapman and Hall, London.

Cook, R. D., Weisberg, S., 1999. Applied Regression Including Computing and Graphics. John Wiley and Sons, Inc., New York.

Debnath, A. K., Debnath, G., Shusterman, A. J., Hansch, C., 1992. 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. 19, 37-52.

Eldred, D. V., Jurs, P. C., 1999. Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. SAR QSAR Environ. Res. 10, 75-99.

Eldred, D. V., Weikel, C. L., Jurs, P. C., Kaiser, K. L. E., 1999. Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. Chem. Res. Toxicol. 12, 670-678.

Filip, P. A., Balaban, T. S., Balaban, A. T., 1987. A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlational ability. J. Math. Chem. 1, 61-83.

Frank, I. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35, 109-135.

Frierson, M. R., Klopman, G., Rosenkranz, H. S., 1986. Structure-activity relationships (SARs) among mutagens and carcinogens: a review. Environmental Mutagenesis 8, 283-327.

Golub, G. H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215-223.

Hall, L. H., Mohney, B., Kier, L. B., 1991. The electrotopological state: Structure information at the atomic level for molecular graphs. J. Chem. Inf. Comput. Sci. 31, 76-82.

Hansch, C., Leo, A., Hoekman, D., 1995. Exploring QSAR: Hydrophobic, electronic, and steric constants. American Chemical Society, Washington, D.C.

Hawkins, D. M., Yin, X., 2002. A faster algorithm for ridge regression of reduced rank data. Computational Statistics and Data Analysis 40, 253-262.

Hawkins, D. M., Young, S. S., Rusinko, A., 1997. Analysis of a large structure-activity data set using recursive partitioning. Quant. Struct.-Act. Relat. 16, 296-302.

Katritzky, A. R., Tatham, D. B., 2001. Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. J. Chem. Inf. Comput. Sci. 41, 1162-1176.

Kier, L. B., Hall, L. H., 1986. Molecular Connectivity in Structure-Activity Analysis. Research Studies Press, Letchworth, Hertfordshire, U.K.

Kier, L. B., Hall, L. H., 1999a. The kappa indices for modeling molecular shape and flexibility. In: Devillers, J., Balaban, A. T. (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR Gordon and Breach Science Publishers, Amsterdam, pp. 455-489.

Kier, L. B., Hall, L. H., 1999b. Molecular Structure Description: The Electrotopological State. Academic Press, San Diego, CA.

Kier, L. B., Hall, L. H., Frazer, J. W., 1991. An index of electrotopological state for atoms in molecules. J. Math. Chem. 7, 229-241.

Kier, L. B., Murray, W. J., Randic, M., Hall, L. H., 1976. Molecular connectivity. V. Connectivity series concept applied to diversity. J. Pharm. Sci. 65, 1226-1230.

Li, K.-C., Duan, N., 1989. Regression analysis under link violation. Annals of Statistics 17, 1009-1052.

Molconn-Z. 2000. v 3.50, Hall Associates Consulting Quincy, MA.

Moyer, S. R., Jurs, P. C., 1990. An SAR study of the mutagenicity of PAH compounds in Salmonella typhimurium. Progress in Clinical and Biological Research 340B, 1-10.

Olive, D. J., 1998. Applied Robust Statistics, Ph.D. thesis, University of Minnesota.

Randic, M., 1975. On characterization of molecular branching. J. Am. Chem. Soc. 97, 6609-6615.

Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B., Basak, S. C., 1984. Discrimination of isomeric structures using information theoretic topological indices. J. Comput. Chem. 5, 581-588.

Serra, R., Jurs, P. C., Kaiser, K. L. E., 2001. Linear regression and computational neural network prediction of *Tetrahymena* acute toxicity of organic compounds from molecular structure. Chem. Res. Toxicol. 14, 1535-1545.

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

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

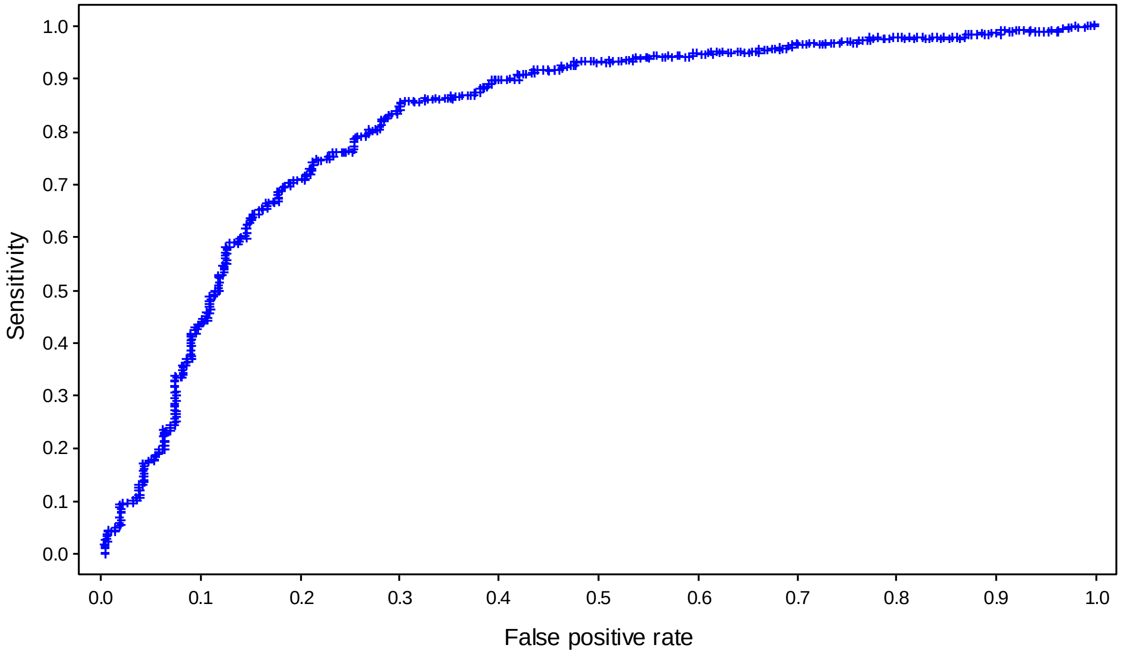
Tripos Associates, Inc. Sybyl Version 6.2; St. Louis, MO, 1995.

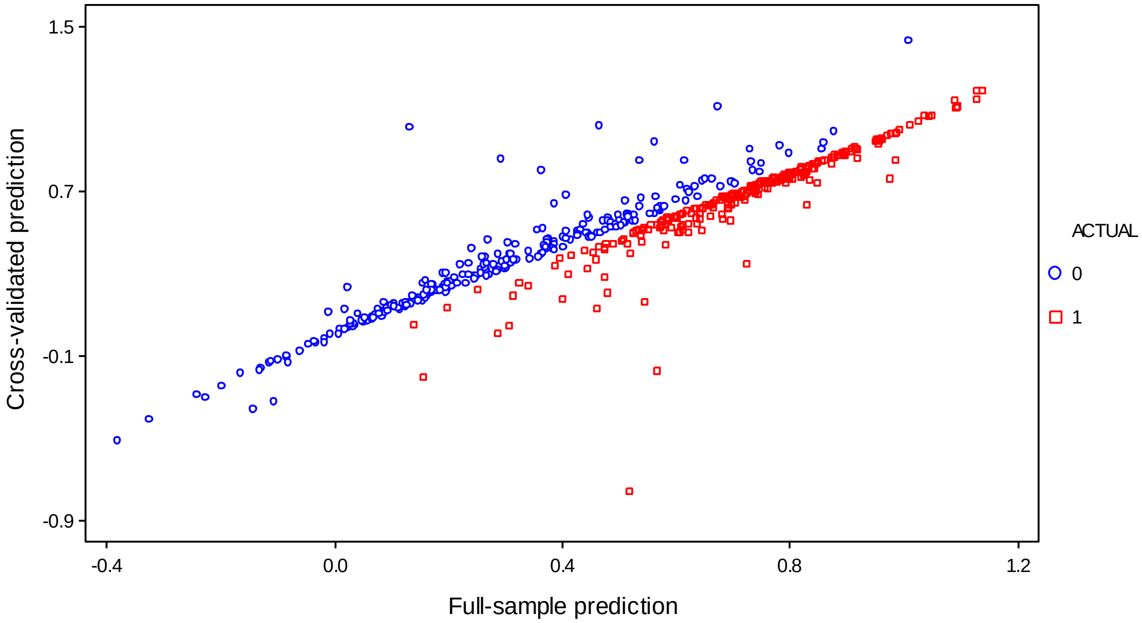
Figure Captions

Figure 1: Receiver operating characteristic (ROC) curve

Figure 2: Fitted-fitted (FF) plot of full-sample and cross-validation predictions

Figure 3: Cook's distance vs. difference between full-sample and cross-validation prediction


Sensitivity
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
False positive rate


1.5
0.7
-0.1
-0.9
Cross-validated prediction
-0.4
0.0
0.4
0.8
1.2
Full-sample prediction
ACTUAL
○ 0
□ 1

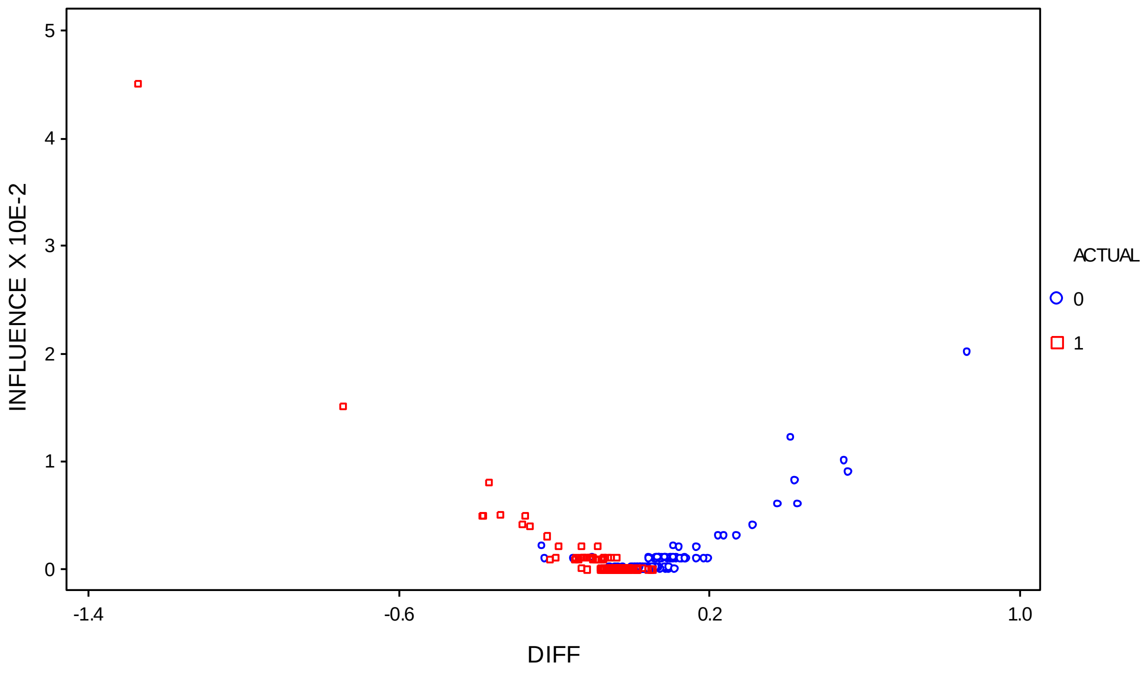

INFLUENCE X 10E-2
5
4
3
2
1
0
-1.4
-0.6
0.2
1.0
DIFF
ACTUAL
○ 0
□ 1

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

Table 2. Compounds that are highly influential in the model fitting.

No.	Chemical name	CAS No.	Influence
1	Endosulfan (TG)	115-29-7	0.045
2	Magenta/Pararosaniline	569-61-9	0.020
3	1,1-Diphenyl-2-butynylcarbamate	20930-10-3	0.015
4	Dinitrosopentamethylenetetramine	101-25-7	0.012
5	Chlorothalonil (TG)	1897-45-6	0.010
6	Trifluraline (TG)	1582-09-8	0.009
7	6-Mercaptopurine	50-44-2	0.008
8	Hexachlorocyclohexane (Lindane)	58-89-9	0.008
9	Caffeine	58-08-2	0.006
10	Methapyrilene	91-80-5	0.006
11	Lasiocarpine	303-34-4	0.005
12	Tris (2,3-Dibromopropyl) phosphate	126-72-7	0.005
13	Thiram	137-26-8	0.005
14	N-Nitrosofolic acid	29291-35-8	0.005
15	Streptozotocin	18883-66-4	0.004
16	Rhodamine B	81-88-9	0.004
17	Cycasin	14901-08-7	0.004
18	N-Methotrexate	59-05-2	0.003
19	Dimethyl sulfoxide	67-68-5	0.003
20	Azathioprine	446-86-6	0.003
21	Eosin	15086-94-9	0.003