

Use of Statistical and Neural Net Methods in Predicting Toxicity of Chemicals: A Hierarchical QSAR Approach

Subhash C. Basak

Brian D. Gute

Gregory D. Grunwald

Natural Resources Research Inst.

University of Minnesota

Duluth, MN 55811 (USA)

{sbasak, bgute, ggrunwal}@wyle.nrri.umn.edu

David W. Opitz

Dept. of Computer Science

University of Montana

Missoula, MT 59812 (USA)

opitz@cs.umt.edu

Krishnan Balasubramanian

Dept. of Chem. and Biochem.

Arizona State University

Physical Sciences Bldg., D-106

Tempe, AZ 85287-1604 (USA)

KBalu@asu.edu

Abstract

A contemporary trend in computational toxicology is the prediction of toxicity endpoints and toxic modes of action of chemicals from parameters that can be calculated directly from their molecular structure. Topological, geometrical, substructural, and quantum chemical parameters fall into this category. We have been involved in the development of a new hierarchical quantitative structure-activity relationship (QSAR) approach in predicting physicochemical, biomedical and toxicological properties of various sets of chemicals. This approach uses increasingly more complex molecular descriptors for model building in a graduated manner. In this paper we will apply statistical and neural net methods in the development of QSAR models for predicting toxicity of chemicals using topostructural, topochemical, geometrical, and quantum chemical indices. The utility and limitations of the approach will be discussed.

Introduction

In 1998 the number of chemicals registered with the Chemical Abstract Service (CAS) rose to over 19 million (CAS 1999). This is an increase of over 3 million chemicals between 1996 and 1998. It would certainly be desirable to be able to test each of these chemicals for their effects on the environment and human health (which we refer to as *hazard assessment*); however, completing the battery of tests necessary for the proper hazard assessment of even a single compound is a costly and time-consuming process. Therefore, there is simply not enough time or money to complete these test batteries for even a tiny portion of the compounds which are registered today (Menzel 1995). An alternative to these traditional test batteries is to develop computational models for hazard assessment. Computational models are fast (milliseconds per compound), cheap (less than one cent per compound), and do not run the risk of adversely affecting the environment during testing. Thus computational models can easily process *all* registered chemicals and flag the ones that require further testing. The central problem with this approach is developing class specific models that can be considered accurate

enough to be useful. In this paper, we present computational models for hazard assessment that are indeed considered both accurate and useful.

One of the fundamental principles of biochemistry is that activity is dictated by structure (Hansch 1976). Following this principle, one can use theoretical molecular descriptors that quantify structural aspects of a molecule to quantitatively determine its activity (Basak & Grunwald 1995; Cramer, Famini, & Lowrey 1993). These theoretical descriptors can be generated directly from the known structure of the molecule and used to estimate its properties, without the need for further experimental data. This is important due to the fact that, with chemicals needing to be evaluated for hazard assessment, there is a scarcity of available experimental data that is normally required as inputs (i.e., independent variables) to traditional quantitative structure-activity relationship (QSAR) model development. A QSAR model based solely on theoretical descriptors on the other hand can process all registered chemicals for hazard assessment. Our recent studies show that hierarchical QSARs (H-QSAR) using theoretical structural descriptors give reasonable models for predicting toxicity (Basak, Gute, & Grunwald In press; Gute & Basak 1997; Gute, Grunwald, & Basak In press).

One potential problem with using our hierarchical approach is that it often gives many independent variables as compared to data points. For instance, in our case study of predicting acute toxicity (LC_{50}) of benzene derivatives, we have 95 independent variables and 69 data points. Therefore, reducing the number of independent variables is critical when attempting to model small data sets. The smaller the data set, the greater the chance of spurious error when using a large number of independent variables (descriptors). Part of our focus in this paper is attempting to reduce the size of the data set.

Hierarchical QSAR

Our recent studies have focused on the role of different classes of theoretical descriptors of increasing lev-

els of complexity and their utility in QSAR (Gute & Basak 1997; Gute, Grunwald, & Basak In press). Four distinct sets of theoretical descriptors have been used in this study: topostructural, topochemical, geometric, and quantum chemical indices. Gute and Basak 1997 provide the detailed list of the indices included in our study.

Topological Indices

The complete set of topological indices used in this study, both the topostructural and the topochemical, have been calculated using POLLY 2.3 (Basak, Harriss, & Magnuson 1988) and software developed by the authors. These indices include the Wiener index (Wiener 1947), the connectivity indices developed by Randic 1975 and higher order connectivity indices formulated by Kier and Hall 1986, bonding connectivity indices defined by Basak and Magnuson 1988, a set of information theoretic indices defined on the distance matrices of simple molecular graphs (Hansch & Leo 1995), and neighborhood complexity indices of hydrogen-filled molecular graphs, and Balaban's 1983 J indices.

Geometrical Indices

The geometrical indices are three-dimensional Wiener numbers for hydrogen-filled molecular structure, hydrogen-suppressed molecular structure, and van der Waals volume. Van der Waals volume, V_W (Bondi 1964), was calculated using Sybyl 6.1 from Tripos Associates, Inc. of St. Louis. The 3-D Wiener numbers were calculated by Sybyl using an SPL (Sybyl Programming Language) program developed in our lab (SYBYL 1998). Calculation of 3-D Wiener numbers consists of the sum entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates for the atoms were determined using CONCORD 3.0.1 from Tripos Associates, Inc. Two variants of the 3-D Wiener number were calculated: ${}^3D W_H$ and ${}^3D W$. For ${}^3D W_H$, hydrogen atoms are included in the computations and for ${}^3D W$ hydrogen atoms are excluded from the computations.

Quantum Chemical Parameters

The following quantum chemical parameters were calculated using the Austin Model version one (AM1) semi-empirical Hamiltonian: energy of the highest occupied molecular orbital (E_{HOMO}), energy of the second highest occupied molecular orbital (E_{HOMO1}), energy of the lowest unoccupied molecular orbital (E_{LUMO}), energy of the second lowest unoccupied molecular orbital (E_{LUMO1}), heat of formation (ΔH_f), and dipole moment (μ). These parameters were calculated using MOPAC 6.00 in the SYBYL interface (Stewart 1990).

Results

We tested the utility of our approach of generating numerous hierarchical theoretical descriptors of com-

pounds on the acute aquatic toxicity (LC_{50}) of a congeneric set of 69 benzene derivatives. The data was taken from the work of Hall, Kier and Phipps 1984 where acute aquatic toxicity was measured in fathead minnow (*Pimephales promelas*). Their data was compiled from eight other sources, as well as some original work which was conducted at the U.S. Environmental Protection Agency (USEPA) Environmental Research Laboratory in Duluth, Minnesota. This set of chemicals was composed of benzene and 68 substituted benzene derivatives. According to the authors, these benzene derivatives were tested using methodologies comparable to their own 96-hour fathead minnow toxicity test system. The derivatives chosen for this study have seven different substituent groups that are present in at least six of the molecules. These groups consist of chloro, bromo, nitro, methyl, methoxyl, hydroxyl, and amino substituents.

We studied two classes of approaches for modeling toxicity: (1) giving all the descriptors to a learning algorithm (neural networks in this case), and (2) reducing the feature set before giving the (reduced) feature set to a learning algorithm. Results for our approaches are from leave-one-out experiments (i.e., 69 training/test set partitions). Leave-one-out works by leaving one data point out of the training set and giving the remaining instances (68 in this case) to the learning algorithms for training. (It is worth noting that each member of the ensemble sees the same 68 training instances for each training/test set partition and thus ensembles have no unfair advantage over other learners.) This process is repeated 69 times so that each example is a part of the test set once and only once. Leave-one-out tests *generalization* accuracy of a learner, whereas training set accuracy tests only the learner's ability to memorize. Generalization error from the test set is the true test of accuracy and is what we report here.

Table 1 gives our results. First we trained neural networks using all 95 parameters. The networks contained 15 hidden units and we trained the networks for 1000 epochs. We normalized each input parameter to a values between 0 and 1 before training. Additional parameter settings for the neural networks included a learning rate of 0.05, a momentum term of 0.1, and weights initialized randomly between -0.25 and 0.25. With these ninety-five parameters, the neural network obtained a test-set correlation coefficient between predicted toxicity and measured toxicity (explained variance) of $R^2 = 0.868$ and a standard error of 0.29. Target toxicity measurements ranged from 3.04 to 6.37.

For our next experiments, the VARCLUS method of SAS 1998 was used for selecting subsets of topostructural and topochemical parameters for QSAR model development. With this method, the set of topological indices is first partitioned into two distinct sets, the topostructural indices and the topochemical indices. To further reduce the number of independent variables for model construction, the sets of topostructural and topochemical indices were further divided into subsets,

Table 1: Relative effectiveness of statistical and neural network methods in estimating LC_{50} of 69 benzene derivatives.

Method	R^2	Standard Error
Linear regression	0.825	0.32
NN with 95 inputs	0.868	0.29
NN with VARCLUS	0.878	0.28

or clusters, based on the correlation matrix using the VARCLUS procedure. This procedure divides the set of indices into disjoint clusters, such that each cluster is essentially unidimensional. From each cluster we selected the index most correlated with the cluster, as well as any indices which were poorly correlated with their cluster ($R^2 < 0.70$). The variable clustering and selection of indices was performed independently for both the topostructural and topochemical indices. This procedure resulted in a set of five topostructural indices and a set of nine topochemical indices. These indices were combined with the three geometric and six quantum chemical parameters described earlier.

The linear regression approach was that described earlier by Gute and Basak 1997. This study found that an accurate linear regression model for acute aquatic toxicity required descriptors from all four levels of the hierarchy: topostructural, topochemical, geometrical and quantum chemical. This model utilized seven descriptors and obtained an explained variance (R^2) of 0.863 and a standard error of 0.30. A leave-one-out approach was then implemented to test the predictivity of the model. This testing resulted in a model with an $R^2 = 0.825$ and a standard error of 0.32.

We also trained neural networks using the 23 parameters provided by this data reduction technique. The parameter settings for these networks were the same as the settings for the other neural network experiments mentioned above. With these 23 parameters, the neural networks obtained a test-set explained variance (R^2) of 0.878 and a standard error of 0.28. Thus the inputs selected by our data reduction procedure were able to increase the accuracy of the neural network.

Discussion and Future Work

The results show that both statistical and neural network methods give acceptable estimates of toxicity. The neural network methods produced improvement over the statistical model. While the method proposed here has proven effective, there is much future work that needs to be completed. For example, though our results demonstrate that our method is able to accurately predict toxicity directly from structure, it would be interesting to know just how many compounds are needed to learn an accurate model of toxicity. Future work, then, is to empirically answer this question. We plan to run our techniques on further reduced data sets and plot leave-one-out accuracy. This would allow one to look

at a curve that plots accuracy versus training set size and decide how many compounds need to be explicitly tested for toxicity.

In the machine learning literature, the process of finding and removing the variables that are unhelpful or destructive to learning is called feature selection (Kohavi & John 1997). Previous work on feature selection has focused on finding the appropriate subset of relevant features to be used in constructing *one* inference model, such as our approach presented in this paper; however, it is appropriate to start considering feature selection with regards to ensembles. An "ensemble" is a combination of the outputs from a *set* of models that are generated from separately trained inductive learning algorithms. Ensembles have been shown, in most cases, to greatly improve generalization accuracy over a single learning model (Breiman 1996a; Maclin & Opitz 1997; Opitz & Shavlik 1996b; Shapire *et al.* 1997). Recent research has shown that an effective ensemble should consist of a set of models that are not only highly correct, but ones that make their errors on different parts of the input space as well (Hansen & Salamon 1990; Krogh & Vedelsby 1995; Opitz & Shavlik 1996a).

Varying the feature subsets used by each member of the ensemble helps promote the necessary diversity and create a more effective ensemble (Opitz submitted). Thus, this concept is particularly appropriate for large feature sets of partially correlated inputs, such as found in hazard assessment of compounds. Ensemble feature selection algorithms, then, not only have the traditional feature-selection criteria of needing to find feature subsets that are germane to the particular task and inductive-learning algorithm, but have the additional criterion of finding a *set* of features subsets that will promote disagreement among the component members of the ensemble.

The ensemble techniques we plan to test are analogous to the popular and successful ensemble approach Bagging (Breiman 1996b). Bagging is a statistical "boot-strap" (Efron & Tibshirani 1993) ensemble method that creates individuals for its ensemble by training each predictor on a random redistribution of the training set. Each predictor's training set is generated by randomly drawing, with replacement, N examples – where N is the size of the original training set; many of the original examples may be repeated in the resulting training set while others may be left out. Each individual predictor in the ensemble is generated with a different random sampling of the training set. Breiman 1996a showed that Bagging is effective on "unstable" learning algorithms where small changes in the training set result in large changes in predictions. This shows that, on average, more diversity is created among the predictors by varying our training set in this manner than is lost in individual predictor accuracy by not training each predictor on the whole data set.

Bagging is not appropriate for most toxicity domains since they are data poor and one cannot afford to waste training examples; however, these domains are feature

rich and thus we can attempt to create diversity by instead varying the inputs to the learning algorithms. Thus we plan to test the approach where each predictor's feature set is generated by randomly drawing, with replacement, N features – where N is the size of the original feature set.

Acknowledgements

This work was partially supported by National Science Foundation grant IRI-9734419, a University of Montana MONTS grant, and U.S. Air Force grants F49620-94-1-0401 and F49620-96-1-0330.

References

- Balaban, A. 1983. Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.* 55:199–206.
- Basak, S., and Grunwald, G. 1995. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry* 19:231–237.
- Basak, S., and Magnuson, V. 1988. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* 19:17–44.
- Basak, S.; Gute, B.; and Grunwald, G. In press. A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In Devillers, J., and Balaban, A., eds., *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach.
- Basak, S.; Harriss, D.; and Magnuson, V. 1988. Polly 2.3. Copyright of the University of Minnesota.
- Bondi, A. 1964. Van der waals volumes and radii. *J. Phys. Chem.* 68:441–451.
- Breiman, L. 1996a. Bagging predictors. *Machine Learning* 24(2):123–140.
- Breiman, L. 1996b. Stacked regressions. *Machine Learning* 24(1):49–64.
- CAS. 1999. The latest cas registry number and substance count. <http://www.cas.org/cgi-bin/regreport.pl>.
- Cramer, C.; Famini, G.; and Lowrey, A. 1993. Use of calculated quantum chemical properties as surrogates for solvatochromic parameters in structure-activity relationships. *Acc. Chemical Research* 26:599–605.
- Efron, B., and Tibshirani, R. 1993. *An introduction to the Bootstrap*. New York: Chapman and Hall.
- Gute, B., and Basak, S. 1997. Predicting acute toxicity (LC50) of benzen derivatives using theoretical molecular descriptors: A hierarchical QSAR approach. *SAR and QSAR in Environmental Research* 7:117–131.
- Gute, B.; Grunwald, G.; and Basak, S. In press. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. In *SAR and QSAR in Environmental Research*.
- Hall, L.; Kier, L.; and Phipps, G. 1984. Structure-activity relationship studies on the toxicities of benzene derivatives: I. an additivity model. *Environ. Toxicol. Chem.* 3:355–365.
- Hansch, C., and Leo, A. 1995. Exploring QSAR: Fundamentals and applications in chemistry and biology. *American Chemical Society* 557.
- Hansch, C. 1976. On the structure of medicinal chemistry. *Journal of Medicinal Chemistry* 19:1–6.
- Hansen, L., and Salamon, P. 1990. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.
- Kier, L., and Hall, L. 1986. *Molecular Connectivity in Structure-Activity Analysis*. Hertfordshire, UK: Research Studies Press.
- Kohavi, F., and John, G. 1997. Wrappers for feature subset selection. *Artificial Intelligence*.
- Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems*, volume 7, 231–238.
- Maclin, R., and Opitz, D. 1997. An empirical evaluation of bagging and boosting. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 546–551.
- Menzel, D. 1995. Extrapolating the future: research trends in modeling. *Toxicology Letters* 79:299–303.
- Opitz, D., and Shavlik, J. 1996a. Actively searching for an effective neural-network ensemble. *Connection Science* 8(3/4):337–353.
- Opitz, D., and Shavlik, J. 1996b. Generating accurate and diverse members of a neural-network ensemble. In Touretsky, D.; Mozer, M.; and Hasselmo, M., eds., *Advances in Neural Information Processing Systems*, volume 8. Cambridge, MA: MIT Press.
- Opitz, D. submitted. Feature selection for ensembles. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*.
- Randic, M. 1975. On characterization of molecular branching. *Journal of American Chemical Society* 97:6609–6615.
- SAS. 1998. Cary, NC: SAS Institute Inc. chapter SAS/STAT User's Guide, Release 6.03 Edition.
- Shapire, R.; Freund, Y.; Bartlett, P.; and Lee, W. 1997. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the Fourteenth International Conference on Machine Learning*, 322–330. Nashville, TN: Morgan Kaufmann.
- Stewart, J. 1990. Mopac version 6.00. qcpe #455. US Air Force Academy, CO: Frank J. Seiler Research Laboratory.
- SYBYL. 1998. Sybyl version 6.1. Tripos Associates, Inc.
- Wiener, H. 1947. Structural determination of paraffin boiling points. *Journal of Am. Chem. Soc.* 69:17–20.