Simulation of the ^{13}C Nuclear Magnetic Resonance Spectra of Ribonucleosides Using Multiple Linear Regression Analysis and Neural Networks

Deborah L. Clouser and Peter C. Jurs*

152 Davey Laboratory, The Pennsylvania State University, University Park, Pennsylvania 16802

Received June 13, 1995[⊗]

Regression equations have been developed to predict the ^{13}C NMR spectra of 17 ribonucleosides through the use of atomic environmental descriptors. These descriptors were calculated directly from the structure of the compounds. Fifteen compounds are used as a training set for linear regression analysis, and two compounds are used as an external prediction set. Due to the diverse nature of the atoms within the data set, the chemical shifts were divided into subsets. The results for each subset are reported. Computational neural networks are also used to predict the chemical shifts of the atoms in the subsets.

INTRODUCTION

Ribonucleosides are a set of biological compounds made up of a nucleic acid and ribose. In this work, the nucleic acid is pure or pyrrolo[2,3-*d*]pyrimidine. ^{13}C NMR has become a valuable tool for investigating the structure of these compounds. In this work, the spectra of some ribonucleosides are simulated and compared to their observed spectra.

One technique used for calculating the chemical shifts of various organic compounds is empirical modeling. This technique makes use of linear models derived from a set of atom-based descriptors, which are numerical representations of the environment surrounding the atoms, calculated for a set of compounds whose chemical shifts are known. Linear regression relates the calculated chemical shift of an atom to the descriptors by the following equation:

$$S = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

where S is the calculated chemical shift, β_n is the coefficient as determined by linear regression, and X_n is the value of the atom-based descriptor. After models are developed, they can be used to predict the chemical shifts of atoms not used during regression analysis.

Another method used for calculating chemical shifts is computational neural networks. This work utilizes fully connected, feed-forward networks utilizing a quasi-Newton training algorithm. This algorithm is based on the works of Broyden,[1] Fletcher,[2,3] Goldfarb,[4] and Shanno.[5] The computational neural networks used here have one output neuron, which produces a value for the chemical shift. The function of these networks is similar to nonlinear regression. The inputs for these networks are the descriptors that were found to be important as determined by multiple linear regression, making these networks model-free, as a model need not be chosen before training begins.

EXPERIMENTAL SECTION

The ^{13}C NMR chemical shifts for the 17 ribonucleosides were found in the literature.[6] The software used for this work is incorporated as part of the ADAPT software package,[7,8] which is installed on a Sun 4/110 workstation in operation at the Pennsylvania State University. The quasi-Newton algorithm is installed on a DEC 3000 AXP 500 workstation.

Data Set Characterization. The 17 compounds used in this study are shown in Table 1. Compounds **1**−**15** were used as a training set. These are the compounds for which the linear regression equations were developed. Compounds **16** and **17** were used as an external prediction set. These compounds were used to test the external predictive ability of the models. The compounds that are present in this external prediction set were chosen randomly as this is the most statistically sound method for choosing a prediction set.

Structure Entry and Modeling. Templates were used to enter the compounds into the computer. These templates generated two-dimensional coordinates which were then read into ADAPT. These structures were then modeled into energy-minimized, three-dimensional coordinates using MOPAC.[9]

Unique Carbon Atom Perception. Once the compounds have been modeled into reasonable three-dimensional coordinates, the unique carbon atoms in each compound must be identified. A unique carbon atom is one that is not chemically equivalent to any other carbon atom in a compound. Only the unique carbon atoms are predicted, as this prevents any undue bias towards a particular atom type. Compounds **1**−**15** contained 181 unique carbon atoms, while compounds **16** and **17** contained 21 unique carbon atoms.

Descriptor Calculation. After the unique carbon atoms have been identified, the atomic environment surrounding each carbon atom is encoded through the use of atom-centered descriptors. These descriptors encode the topological, geometrical, and electronic features of the structural surroundings of the carbon atoms. ADAPT has the capability to calculate over 700 descriptors, although not all descriptors may be applicable to each data set. To ensure that only meaningful and statistically valid descriptors are used to develop regression equations, descriptors containing more than 70% identical or zero values are eliminated. The remaining descriptors are then screened for pairwise correlation. If a pair of descriptors is flagged as having a high correlation, $r > 0.90$, than one descriptor is removed. The descriptor that is the easiest to calculate remains. However, if intuition indicates that the more complicated descriptor may be more useful, it may be retained. After this screening

[⊗] Abstract published in *Advance ACS Abstracts*, November 1, 1995.

0095-2338/96/1636-0168$12.00/0

Table 1. Ribonucleosides Used in This Study

1	4-amino-7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine
2	4-amino-5-cyano-7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine
3	7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidin-4-one
4	7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine-4-thione
5	9-(β-D-ribofuranosyl)purine
6	7-(β-D-ribofuranosyl)adenine
7	7-(β-D-ribofuranosyl)hypoxanthine
8	1-methyl-9-(β-D-ribofuranosyl)hypoxanthine
9	6-methoxy-9-(β-D-ribofuranosyl)purine
10	9-(β-D-ribofuranosyl)purine-6-thione
11	1-methyl-9-(β-D-ribofuranosyl)purine-6-thione
12	6-methylthio-9-(β-D-ribofuranosyl)purine
13	4-amino-5-carboxamido-7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine
14	9-(β-D-ribofuranosyl)adenine
15	7-(β-D-ribofuranosyl)purine-6-thione
16	7-(β-D-ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine
17	9-(β-D-ribofuranosyl)hypoxanthine

process, more than 180 topological, geometrical, and electronic descriptors remained in the pool to be used for regression analysis.

Descriptor Selection. Given a suitable pool of descriptors, models can be developed. It would be far too time-consuming a process to test every combination of descriptors possible, so a method of selecting optimum subsets of descriptors needs to be used. The method utilized in this study makes use of a generalized simulated annealing (GSA) algorithm[10,11] which is based on the physical process of annealing a solid metal. This algorithm selects a random subset of descriptors from the pool being used to form a model, and then other descriptors are added to an deleted from this model, depending on their performance in the model. The criterion used for this process is the rms error of the prediction. The number of descriptors included in a model is determined by the user. Unlike other methods, GSA will accept several detrimental steps, allowing it to work out of a local minimum. This increases the likelihood of finding a global minimum, even though there is no check to determine whether a global minimum has been found.

RESULTS AND DISCUSSION

Multiple Linear Regression Analysis. In the beginning of this study, all the unique carbon atoms were in one set, and models were being developed for this entire set. It was discovered during the model-building process that the types of carbon atoms present in the ribonucleosides encompassed too broad a range to predict as one set. The models developed for this entire set were very poor, with high rms errors and low R values.

Therefore, the training set was broken down into two subsets in an attempt to develop models with more accurate predictive capabilities. The first subset, which contained 79 atoms, consisted of all of the unique carbon atoms contained in the ribofuranoses. The second subset, which contained 82 atoms, consisted of all of the unique carbon atoms present in the nucleosides. The external prediction set was also split in this manner, so that subset 1 had 10 prediction set atoms and subset 2 had 11 prediction set atoms.

Before developing models for these subsets, the descriptors were screened again. This was done in order to eliminate any statistical problems that may have arisen due to the data being grouped in more homogeneous subsets. The descriptors are calculated for the entire data set at once, and when the data set is broken down into subsets, so are the descriptors. This can create correlation problems that were not present for the data set as a whole, and some descriptors may contain identical values and pairwise correlations for the new subsets. Descriptor screening ensures that information-rich descriptors are used for the model-building process for the subsets. After this process was completed, the pools of descriptors that remained for each subset were smaller than the pool for the entire data set. The subset pools each contained about 100 topological, geometrical, and electronic descriptors.

Using GSA, a four-descriptor model was found to give very reasonable results for subset 1, the atoms in the ribofuranose rings. The statistics for this model are shown in the top section of Table 2. Descriptors AVC1, AVC3 and WPAT are topological in nature while AVBS is geometrical. This group of descriptors is capturing the effects that were important in determining the chemical shifts of these subsets of atoms. The training set error for this model was 0.88 ppm with an external prediction set error of 0.48 ppm. The fact that the calculated error was lower for the prediction set than for the training set is simply a function of the atoms that are contained in the prediction set. A plot of the calculated chemical shifts versus the observed chemical shifts is shown in Figure 1. The shift at 11 ppm looks as though it may be a shift that has a high leverage effect on the line, but internal validation showed that for this model the shift at 11 ppm was not biasing the regression equation. This validation involved removing the shift at 11 ppm, performing the regression using the same four descriptors shown above, and checking the coefficients of the linear regression. Since the coefficients did not change, the shift is not biasing the regression equation.

Using GSA, a nine-descriptor model was then found for subset 2, the atoms contained in the nucleosides. A larger set of descriptors was needed to capture the influencing factors that determine the chemical shifts of this subset of atoms as they were in a more complicated chemical environment than subset 1. The training set error for this model was 3.29 ppm, and the external prediction set error was 3.10 ppm. The statistics for this model are shown in the bottom section of Table 2. AVC3, ACNC, and TESV are topological, AVCG, MPCG, NTCH, and TOHC are electronic, and CO8D and HND3 are geometrical. AVC3 is also present in the model developed for subset 1. The errors for this model are well above the standard of 1.0 ppm, which is the goal of work such as this. The reason for this is there are

Table 2. Regression Models Developed during This Study

descriptor[a]	mean	SD[b]	coefficient	mean effect[c] (ppm)
		Model Developed for Atoms Contained in Subset 1		
AVC1 4	0.608	0.564	2.76 ± 0.20	1.68 ± 0.12
AVC3 1	1.54	0.550	−10.3 ± 0.26	−15.9 ± 0.40
WPAT 1	3.15	0.277	58.4 ± 0.50	184 ± 1.58
AVBS 1	1.5	0.0387	−111 ± 2.8	−166 ± 4.2
intercept			71.1 ± 4.4	
		$n = 79$ $s = 0.88$ ppm $R = 0.998$		
		$n = 10$ $s = 0.48$ ppm $R = 0.999$		
		Model Developed for Atoms Contained within Subset 2		
AVC3 1	1.28	0.810	10.4 ± 1.0	13.3 ± 1.2
ACNC 1	0.262	0.0365	174 ± 12	45.5 ± 3.1
TESV 1	0.696	0.651	−10.2 ± 1.1	−7.10 ± 0.76
AVCG 1	−0.0997	0.107	−42.9 ± 9.4	4.29 ± 0.94
MPCG 1	0.0192	0.129	102 ± 12	1.96 ± 0.23
NTCH 1	−0.781	0.955	−21.7 ± 0.90	16.9 ± 0.70
TOHC 3	2.24	0.848	−4.02 ± 0.60	−9.00 ± 1.3
CO8D 2	0.0637	0.0899	24.3 ± 4.8	1.55 ± 0.31
HND3 1	0.0690	0.0931	153 ± 14	10.6 ± 0.97
intercept			62.6 ± 4.2	
		$n = 82$ $s = 3.29$ ppm $R = 0.985$		
		$n = 11$ $s = 3.10$ ppm $R = 0.986$		

[a] Descriptor definition ("heavy atom/r denotes all non-hydrogen atoms). AVC1 4, the number of primary heavy atoms four bonds from the carbon center; AVC3 1, the number of tertiary heavy atoms one bond from the carbon center; WPAT 1, the sum of the weighted paths originating from the carbon center; AVBS 1, the average length of the bonds attached to the carbon center; ACNC 1, the average connectivity index over bonds one bond away from the carbon center; TESV 1, the electropological state of the carbon center; AVCG 1, the average σ charge for all heavy atoms one bond from the carbon center, MPCG 1, the most positive σ charge among heavy atoms one bond away; NTHC 1, the sum of the extended Hückel charges for all heavy atoms one bond from the carbon center; TOHC 3, the sum of the absolute values of the extended Hückel charges on all heavy atoms three bonds from the carbon center; CO8D 2, the inverse throughspace distance squared from the carbon center to carbons with two heavy atoms connected with two aromatic bonds; HND3 1, sum of the inverse throughspace distance from the hydrogens attached to the carbon center to all nitrogens located one bond away. [b] SD, standard deviation. [c] Mean effect, the average shielding and deshielding contribution of each descriptor on the predicted chemical shift.

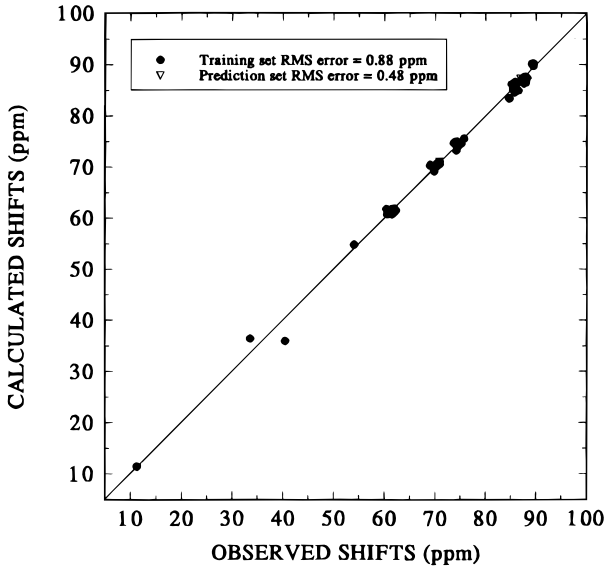


Figure 1. Calculated versus observed shifts from regression for the atoms contained in subset 1.

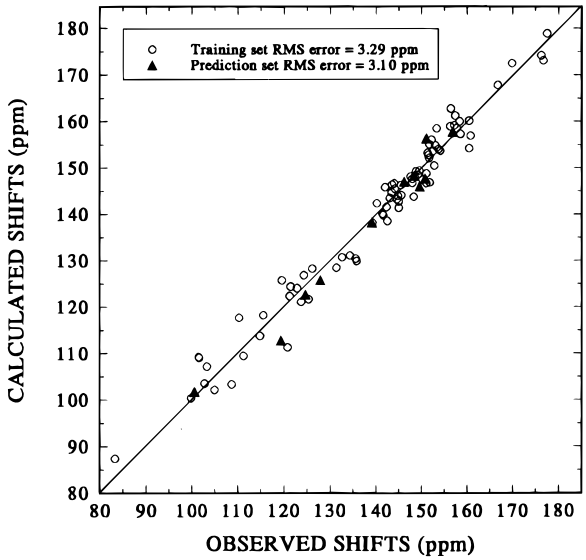


Figure 2. Calculated versus observed shifts from regression for the atoms contained in subset 2.

shifts that are predicted very poorly. However, no predicted shifts were flagged as outliers using statistical tests. The training set error is also higher than the prediction set error due to these poorly predicted shifts. When these shifts are removed, the training set error becomes lower than the prediction set error. However, the goal of this work is to predict all of the shifts for each compound, so these shifts were left in the training set. A plot of the calculated versus the observed shifts is shown in Figure 2. Many regression models were developed for this data set, and this model was one of the best found.

When this data set was investigated using environmental descriptors, nine carbon atoms seemed to form a subset slightly different from the two already described, even though these atoms were contained with the nucleosides, which is subset 2. Their surrounding chemical environments were somewhat different from the other carbon atoms contained within the nucleosides. Some of the atoms in this set were among the atoms that were predicted poorly by the nine-descriptor model. These atoms were placed in subset 1 and predicted, but the results were extremely poor.

One measure of the quality of a model is to perform library searching with the predicted shifts. Accurate models will find observed spectra as the most similar to the predicted spectra. A measure of this similarity is a corresponding score, which is the sum of the squared errors between the calculated shifts and the observed shifts. The search algorithm used is based on a nearest-neighbor approach. In order to perform library searching, the predicted shifts from each model had to be combined to form complete spectra. For the two models, nine observed spectra were number one matches with the predicted spectra, five were number two matches, one was a number three match, and two predicted spectra did not have their observed spectra as a top five match. Even though some spectra were number one matches, the scores, which were explained above, were usually over 50.

Computational Neural Networks. After regression analysis has been completed, and models have been found for each subset, computational neural networks can also be used to predict the chemical shifts for ribonucleosides. The neural networks used here are trained with a quasi-Newton method algorithm, which has been explained in detail elsewhere.[12]

The neural networks in this paper are made up of three layers of neurons. The first layer is called the input layer. No calculations are performed in this layer, which is used simply to input the descriptor values from the models found by linear regression and the observed shifts for each atom. The observed shifts are used a target values. The inputs are scaled so that all values fall between 0 and 1, as this is the range of the sigmoidal transfer function used here. The second layer is called the hidden layer. Here, the data are processed and sent on to the final layer, which contains a single output neuron. The output of this layer is the predicted chemical shift. In a fully connected, feed-forward network, each neuron is connected to every neuron in the level below it. Each connection has associated with it an adjustable weight, which determines how much information is transferred from one neuron to the next. After the error between the predicted shift and the observed shift is found, the weights are adjusted to minimize this error. There is more than one method available for adjusting these weights, but the quasi-Newton method has an advantage over these other methods. This method has the ability to use the shape of the error function to determine the step size taken to adjust the weights. This decreases the amount of time take to find the set of weights that will give the minimum rms error. The quasi-Newton algorithm begins with a random sets of weights, and the minimum rms error found depends on these random weights. Therefore, many different starting sets of weights are used to an effort to find the best set to use for prediction. Coupling the faster algorithm with a fast workstation makes the task of testing different random starting weights relatively simple.

Neural networks tend to improve upon the results of linear regression for two reasons. First, neural networks are able to take advantage of nonlinear information that may be contained within a set of inputs. Second, neural networks often have more adjustable parameters than linear regression. To avoid results due to chance, the number of adjustable parameters should be less than half the number of observations being presented to the network.[13]

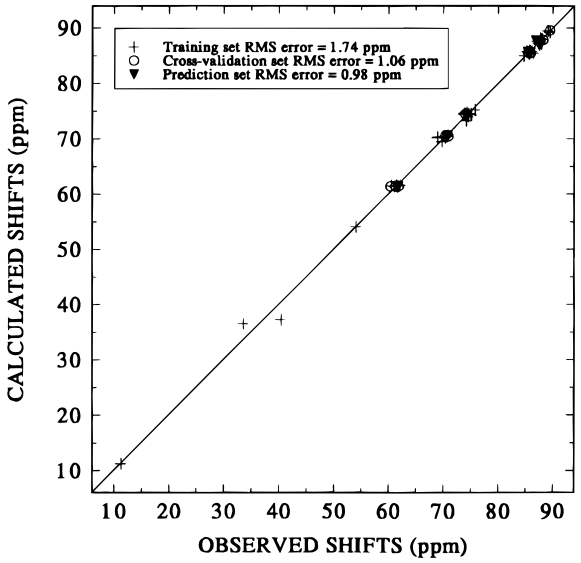


Figure 3. Calculated versus observed shifts from neural networks for subset 1.

When the data for the training set is presented to the network, it must be divided into two sets, a training set and a cross-validation set. The training set contains the carbon atoms being used to train the network, and the cross-validation set is used to prevent overtraining the network. The number of atoms in the cross-validation set is approximately equal to the number of atoms in the external prediction set, which is the same prediction set used for regression analysis. Periodically during training the shifts in the cross-validation set are predicted. When the error of the cross-validation set fails to improve, or starts to increase, then the network is beginning to encode information particular to only that data set, and external predictive ability may be lost. Training is stopped at this point.

The first model submitted to neural networks was the model developed for subset 1. This network has the architecture 4:2:1. The training set contained 69 carbon atoms, the cross-validation set contained 10 atoms, and the external prediction set contained 10 atoms. The training set error for this network was 0.69 ppm, the cross-validation set error was 0.47 ppm, and the external prediction set error was 0.39. A plot of the calculated versus observed shifts is shown in Figure 3. The networks improved slightly as compared to regression, but the regression model itself was very accurate. There are only four descriptors in this model, so there may not be a lot of nonlinear information in this model that can be utilized by neural networks.

The second model submitted to neural networks was the model developed for subset 20. The training set for this network contained 72 carbon atoms, while the cross-validation set contained 10 atoms, and the external prediction set contained 11 atoms. With many networks trained in the beginning of this research, even though improved rms errors were being found for the training set and cross-validation set, the rms error for the prediction set was very large, often in excess of 5 ppm. While searching for the cause of the poor prediction set error, it was noted that compound **2** contained the only cyano carbon atom in the training set, making this type of carbon atom underrepresented in the data set. It may be possible that while the networks were attempting to fit the shift for the cyano carbon, the predictions

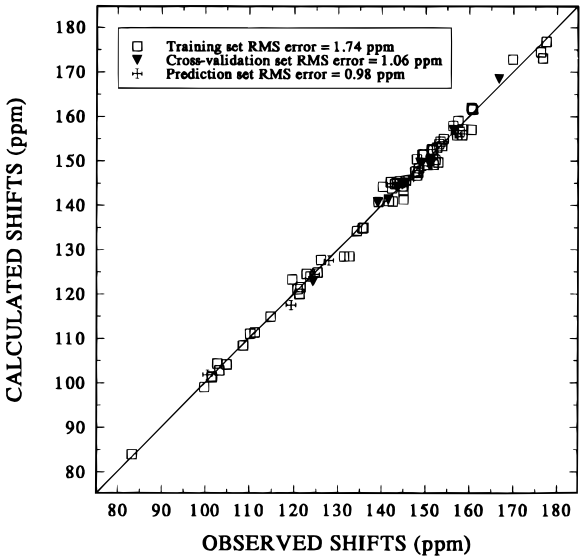
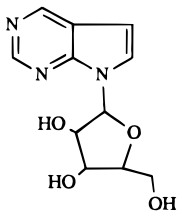


Figure 4. Calculated versus observed shifts from neural networks for subset 2.



7-(β-D-Ribofuranosyl)pyrrolo[2,3-*d*]pyrimidine

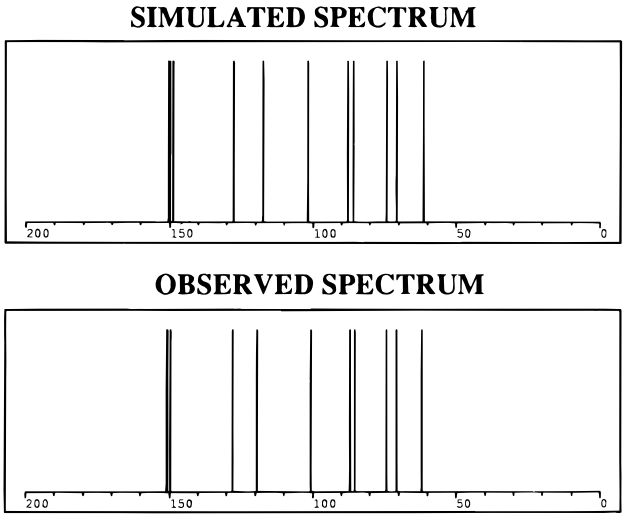


Figure 5. Predicted and observed spectra for compound **16** using computational neural networks.

for other chemical shifts were being adversely affected, thus diminishing the external predictive ability of the network. When the cyano carbon was removed from the training set, the errors for the prediction set greatly improved. The network chosen for this work had a training set error of 1.74 ppm, a cross-validation set error of 1.06 ppm, and an external prediction set error of 0.98 ppm. This error is still above 1.0 ppm, but the neural networks improved substantially upon the error as compared to regression analysis. A plot of the calculated versus observed shifts for this subset is shown in Figure 4. (As a comparison, regression analysis was performed without the cyano carbon in the training set, and the resulting model had an error of 3.19, which is only slightly better than with the cyano carbon present.)

Library searching was also performed on the results of neural networks, and the results were better than those obtained with regression analysis. When the predicted spectra were compared to the observed spectra, all the observed spectra were number one matches. An example of the accuracy of these matches is shown in Figure 5. The spectra shown are for compound **16**, and the rms spectral error for this example is 0.88 ppm.

CONCLUSIONS

Multiple linear regression analysis and computational neural networks were used to predict the ^{13}C NMR shifts of ribonucleosides. Due to the diverse nature of the atom types, the data set was broken down into subsets and investigated. These subsets yielded more accurate results, especially when submitted to neural networks. The results from regression analysis are not as good as was expected for subset 2 due to some poorly predicted shifts, but with the improvement provided by neural networks, the final predictions were acceptable. Accurate predictive equations for biologically active compounds such as ribonucleosides may give researchers a valuable tool to aid in structure elucidation.

REFERENCES AND NOTES

(1) Broyden, C. G. The convergence of a class of double-rank minimization algorithms. *J. Inst. Math. Its Appl.* **1970**, *24*, 76.
(2) Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, *13*, 317.
(3) Fletcher, R. *Practical Methods of Optimization*; Wiley: New York, 1980; Vol. 1.
(4) Goldfarb, D. A. family of variable-metric methods derived by variational means. *Math. Comput.* **1970**, *24*, 23.
(5) Shanno, D. F. Condition of quasi-Newton methods for function minimization. *Math. Comput.* **1970**, *24*, 647.
(6) Chenon, M. T.; Pugmire, R. J.; Grant, D. M.; Panzica, R. P.; Townsend, L. B. A basic set of parameters for the investigation of tautomerism in purines established from carbon-13 magnetic resonance studies using certain purines and pyrrolo[2,3-*d*] pyrimidines. *J. Am. Chem. Soc.* **1975**, *97*, 4627.
(7) Stupper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Functions*; Wiley-Interscience: New York, 1979; pp 83−90.
(8) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olsen, E. C., Christoffersen, R. E., Eds.; American Chemical Society: Washington, DC, 1979; pp 103−129.
(9) Stewart, J. J. P. MOPAC 6.0, *Quantum Chemistry Program Exchange*, Indiana University, Bloomington, IN, 1994; Program 455.
(10) Bohachevsky, I. O.; Johnson, M. E.; Stein, M. L. Generalized simulated annealing for function optimization. *Technometrics* **1986**, *28*, 209−217.
(11) Sutter, J. M.; Jurs, P. C. Automated descriptor selection for quantitative structure−activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(12) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−activity relationships for toxicity for phenols using regression analysis and computational neural networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(13) Livingstone, D. J.; Manallack, D. T. Statistics using neural networks: Chance effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

CI950055Y