Sensitivity of Infrared Spectra to Chemical Functional Groups

ARTICLE *in* ANALYTICAL CHEMISTRY · JULY 2008

Impact Factor: 5.64 · DOI: 10.1021/ac8000429 · Source: PubMed

3 AUTHORS, INCLUDING:

Chris W. Brown
University of Rhode Island
139 PUBLICATIONS **2,280** CITATIONS

Anal. Chem. **2008,** *80,* 4186–4192

Sensitivity of Infrared Spectra to Chemical Functional Groups

Kevin Judge,[†] Chris W. Brown,[†,*] and Lutz Hamel[‡]

Department of Chemistry, and Department of Computer Science and Statistics, University of Rhode Island, Kingston, Rhode Island 02881

Spectral features from specific regions in infrared spectra of organic molecules can consistently be attributed to certain functional groups. Artificial neural networks were employed as a pattern recognition tool to elucidate the relationships between functional groups and spectral features. The ability of these network models to predict the presence and absence of a variety of functional groups was evaluated. The sensitivity of the artificial neural network over the entire infrared spectral region was used to generate a spectral factor representation of the major information associated with each functional group. The resulting sensitivity factors were utilized in a much simpler model for functional group prediction. Ultimately, the presence of a functional group was predicted based on the dot product of an unknown spectrum with the corresponding sensitivity factor. A probability based on Bayes' theorem was assigned to each of the predictions. The prediction accuracies were greater than 90% for all 13 functional groups considered in the investigation.

Vibrational spectroscopy is a fast and reproducible technique for obtaining information relevant to chemical structures.[1] Many chemical functional groups regularly absorb in certain spectral regions, allowing for their identification.[2–6] Early methods for searching and processing infrared spectral libraries relied on manual sorting based on strong absorptions and on correlation charts.[3] With the advent of personal computers in ~1980, library processing and spectral interpretation focused on pattern recognition methodologies[7–10] and expert systems.[11–13] Pattern recognition was initially based on peak positions,[7,8] but later peak widths were used, and eventually principal component analysis (PCA) was added to the processing.[9,10] Initially, expert systems were based on a set of rules for correlating peak positions with the presence of functional groups.[11] Later, Griffiths' group[12,13] improved the expert system methodology by including PCA in the processing. Developing neural networks to interpret infrared spectra has been the goal of numerous investigations since the late 1980s.[14–20] Initial investigations were limited by computer memory and speed, but this rapidly changed with increasing computer power.

In the present study, artificial neural networks were created to analyze spectra for the detection of different functional groups. A related goal was to develop a model to describe the criteria for the determination of the absence or presence of a specific substructure. Lastly, a level of confidence or probability was assigned to the model's results.

Artificial neural networks (ANN) are powerful tools for pattern recognition that have the ability to expose and model nonlinear relationships. Two major drawbacks to using ANNs are their susceptibility to overfitting and a "black box" quality.[21] Overfitting occurs when the network essentially memorizes the training data, losing its ability to generalize. Monitoring the performance of a network using a separate validation set can prevent overfitting. The best model is found when the error in the validation set, not the training set, is at a minimum.

The "black box" characteristic, sometimes referred to as nontransparency, is a result of the complexity of ANNs. As will be discussed in more detail, ANNs contain multiple nodes distributed among several layers, each performing a separate calculation. The network learns by adjusting the weights that connect the nodes of different layers. It is unclear from the weights

* Corresponding author. E-mail: cbrown@chm.uri.edu.
† Department of Chemistry.
‡ Department of Computer Science and Statistics.

(1) Debska, B.; Guzowska-Swider, B. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 325–329.
(2) Bellamy, L. J. *The Infrared Spectra of Complex Molecules*, 2nd ed.; John Wiley & Sons, Inc.: New York, 1958.
(3) Colthup, N. B.; Daly, L. H.; Wiberley, S. E. *Introduction to Infrared and Raman Spectroscopy*; Academic Press, 1990.
(4) Lin-Vien, D.; Colthup N. B.; Fateley, W. G.; Grasselli, J. G. *Infrared and Raman Characteristic Frequencies of Organic Molecules*; Academic Press: New York, 1991.
(5) Yao, J.; Fan, B.; Doucet, J. P. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1046–1052.
(6) Nyquist, R. A. *The Interpretation of Vapor-Phase Infrared Spectra*; Sadtler Research Libraries: Philadelphia, PA, 1984.
(7) Delaney, M. F.; Uden, P. C. *Anal. Chem.* **1979**, *51*, 1242–1249.
(8) Delaney, M. F. *J. Chromatogr. Sci.* **1979**, *17*, 428–43.
(9) Warren, J.; Delaney, M. F. *Appl. Spectrosc.* **1983**, *37*, 172–181.
(10) Domokos, L.; Frank, I.; Matolcsy, G.; Jalsovszky, G. *Anal. Chim. Acta* **1983**, *154*, 181–189.
(11) Tomellini, S. A.; Saperstein, D. D.; Stevenson, J. M.; Smith, G. M.; Woodruff, H. B.; Seelig, P. F. *Anal. Chem.* **1981**, *53*, 2367–2369.
(12) Perkins, J. H.; Hasenoehrl, E. J.; Griffiths, P. R. *Anal. Chem.* **1991**, *63*, 1738–1747.
(13) Hasenoehrl, E. J.; Perkins, J. H.; Griffiths, P. R. *Anal. Chem.* **1992**, *64*, 656–663.
(14) Donahue, S. M. Processing Spectral Data in the Fourier Domain. Ph.D. Thesis, University of Rhode Island, 1988.
(15) Daniel, N. W.; Griffiths, P. R. *Proc. SPIE—Int. Soc. Opt. Eng.* **1993**, *2089*, 230.
(16) Daniel, N. W.; Lewis, I. R.; Griffiths, P. R. *Appl. Spectrosc.* **1997**, *51*, 1868.
(17) Cirovic, D. A. *Trends Anal. Chem.* **1997**, *16*, 148–155.
(18) Penchev, P. N.; Andreev, G. N.; Varmuza, K. *Anal. Chim. Acta* **1999**, *388*, 145–159.
(19) Brown, C. W.; Lo, S. C. *Anal. Chem.* **1998**, *70*, 2983–2990.
(20) Jegla, J. D. Automatic Classification of Organic Compounds from Their Vapor-Phase Infrared Spectra. Ph.D. Thesis, University of Idaho, 1997.
(21) Tu, J. V. *J. Clin. Epidemol.* **1996**, *49*, 1225–1231.

10.1021/ac8000429 CCC: $40.75
Published on Web 04/24/2008

alone what the ANN has learned and how it makes a decision. Through a process known as sensitivity analysis,[21] the most relevant features that the network associates with each decision can be determined. Once these sensitivity factors are obtained, they can be used to construct much simpler models. In these simplified models, each decision is assigned a probability using Bayes' theorem.

The goal of finding a model to relate infrared spectral patterns to specific functional groups dates back to the early investigations on using ANN to interpret spectra. Donahue[14] generated representative patterns and their sensitivity to various functional groups using score plots from PCA on a limited library in the late 1980s. Later, Daniel and Griffiths[15] proposed sensitivity analysis as a means to obtain feature spectra for functional groups, and this method was eventually applied to the identification of nitro explosives.[16] Jegla,[20] working in Griffiths' laboratory, extensively investigated sensitivity analysis on a library of gas-phase spectra. Preprocessing methods including PCA and auto scaling were explored prior to determine feature spectra from sensitivity analysis. The present study takes advantage of the sensitivity analysis method developed by Harrington, et al.[22]

Artificial Neural Networks. Artificial neural networks are models that mimic biological neurons in the brain. In this particular study, a multiple layered perceptron (MLP) network was utilized. The standard MLP network includes three layers, known as the input layer, hidden layer, and output layer.[23–25] The input layer contains a neuron, or node, for each input variable of a sample. The input values, x, are weighted, w, and passed to the hidden layer, which can contain any number of nodes. Each hidden neuron sums these weighted inputs and adds a bias, b_j,

$$\mathrm{net}_j = \left[\sum w_{ij}x_i\right] - b_j \tag{1}$$

A transfer function, $f(\mathrm{net}_j)$, is utilized to calculate a final output. A typical choice for this function is the nonlinear sigmoid function,

$$f(\mathrm{net}_j) = 1/[1 + \mathrm{e}^{-\mathrm{net}j}] \tag{2}$$

which forces the final neuron output to a value between 0 and 1.[15] Training is achieved by modifying the weights to minimize the error between the predicted outputs and the target values using error back-propagation.[16] The transfer functions employed during the analysis were the sigmoid function at the hidden layer and the linear function for the output layer.

The application of two constraints to the network served several purposes.[26] An unconstrained network can reduce its output error by either increasing the length of the weight vector, w_j, or orienting the weight vector, w_j, and bias, b_j, to a more optimal solution. By dividing the weighted sum of inputs by the length of the weight vector

$$\mathrm{net}_j = \left[(1/|w_j|)\sum w_{ij}x_i\right] - b_j \tag{3}$$

the weight vector is normalized and kept at a constant length. This constraint causes the rotation of the weight vector to be the driving force in minimizing the network's error. This was the critical modification, and its importance in realizing spectral relationships will be illustrated later. The sigmoidal, nonlinear transfer function, $f(\mathrm{net}_j)$, was also altered with the introduction of a constraint, t_j,

$$f(\mathrm{net}_j) = 1/[1 + \mathrm{e}^{-(\mathrm{net}j/\mathrm{tj})}] \tag{4}$$

The addition of t_j controlled the steepness of the sigmoid curve. It has been reported that adjusting the steepness of the sigmoid curve does not have much of an impact on the network's performance.[27] However, a very large value of t_j would result in a fairly flat transfer function, making it difficult and time-consuming for the network to grasp the most obvious trends in the data. On the other hand, using a value that is less than 1.0, the transfer function becomes steeper and forces the network to make faster decisions, which reduces the training time and further prevents overfitting.

Sensitivity Analysis. Sensitivity analysis is a method used to extract the features most responsible for the decision of an ANN.[21] It is a measure of the change in the network's response with a change in each input variable, or the gradient of the network output. Considering that ANNs are nonlinear functions, the gradient is dependent upon the point at which it is evaluated. In this study, the mean spectrum, x_{mean}, of the input samples in the training set containing the functional group of interest was used. The sensitivity, S_k, of the kth variable was calculated by examining the network output, $F(x)$, when that variable (wavenumber) was perturbed and the others were kept constant. A row vector, B_k, was generated and consisted of all zero values except for the perturbation, p, in the kth position. This vector was added to and subtracted from the mean spectrum for the functional group after the network was trained. This resulted in the following equation for the sensitivity at the kth wavenumber

$$S_k = [F(x_{\mathrm{mean}} + B_k) - F(x_{\mathrm{mean}} - B_k)]/2p \tag{5}$$

where p, the perturbation, was equal to a percentage of the largest input variable of x_{mean}. Empirically, we found that the optimum value for p was 1% of the maximum absorbance (peak) value in the average spectrum, x_{mean}.

Probability and Bayes' Theorem. Probabilities for the predictions based on the sensitivity factors were obtained employing Bayes theorem,[24] which allows for the determination of the posterior probability of an event h, given data D, based on prior knowledge of the probabilities $P(h)$, $P(D)$, and $P(D|h)$:

$$P(h|D) = P[(D|h)P(h)]/P(D) \tag{6}$$

The initial probability that an event will occur, $P(h)$, is calculated from the number of samples in which h is true, such as the number of spectra in which a functional group is present, divided

(22) Harrington, P. B.; Urbas, A.; Wan, C. *Anal. Chem.* **2000**, *72*, 5004–5013.
(23) Looney, C. G. *Pattern Recognition Using Neural Networks: Theory and Algorithms for Engineers and Scientists*; Oxford University Press, Inc.: New York, 1997.
(24) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.
(25) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature* **1986**, *323*, 533–536.
(26) Harrington, P. B. *Anal. Chem.* **1994**, *66*, 802–807.
(27) Klawun, C.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 69–81.

Table 1. Distributions of the 13 Functional Groups for the IR Spectral Library

	no. of spectra
aromatic	1413
amine	597
carbonyl	1109
C−O	1553
O−H	881
ketone	251
ester	380
aldehyde	136
acid	256
C=C	346
nitro	78
C≡N	135
alcohol	651
total	2752

by the total population, the number of samples in the library. The probability of getting a result D, a dot product in this study, given an event h, $P(D|h)$, was found from the distributions of the dot products between the sensitivity factor and each library spectrum. Both the dot product distributions, one for samples containing a certain functional group and another for samples in which the structure was absent, were normalized to an area of 1.0. Because the normalizing constant, $P(D)$, is in the equations for $P(h|D)$ in both the presence and absence of a functional group, it cancels out when the probabilities are scaled to total probability of 1.0. The final normalized probability for the presence of the functional group of interest becomes

$$P(h_{pres}|D) = P(D|h_{pres}) \cdot P(h_{pres}) / [P(D|h_{pres}) \cdot P(h_{pres}) + P(D|h_{abs}) \cdot P(h_{abs})] \quad (7)$$

where $P(h_{pres})$ and $P(h_{abs})$ are found using the initial population information and $P(D|h_{pres})$ and $P(D|h_{abs})$ are conditional probabilities of obtaining a certain dot product given the presence and absence of the substructure, respectively. The final probability that the structure is absent given the dot product can be found in a similar fashion, or by just subtracting $P(h_{pres}|D)$ from 1. This is acceptable because there are only two possible events, the functional group is either present or absent from the sample's molecular structure.

EXPERIMENTAL SECTION

Library Construction. The infrared spectral library was assembled and labeled by the presence or absence of 13 functional groups. The library contained 2752 infrared spectra from the Aldrich-SensIR ATR-IR library (Sigma-Aldrich, Inc., Milwaukee, WI and SensIR Technologies now Smiths Detection, Danbury, CT). Each spectrum consisted of 1738 spectral intensities representing the absorbance over the region from 4000 to 650 cm^{-1}. The distributions for each functional group are listed in Table 1.

Data Analysis. Data analysis was executed using MatLab version 7.2 (Mathworks, Natick, MA), which included the Neural Network toolbox. All spectra were normalized to a total area of 1.0. The sensitivity spectra are susceptible to spectral noise especially in regions of lower signals (e.g., at high wavenumbers); thus, the spectra were smoothed with a 15 point, fifth-order polynomial using the Savitzky−Golay algorithm.[28] Several different window sizes and polynomials were tested to determine the optimum point spread and polynomial order for smoothing. The target vectors for the functional group assignments consisted of 1's and 0's indicating the presence and absence of a functional group, respectively.

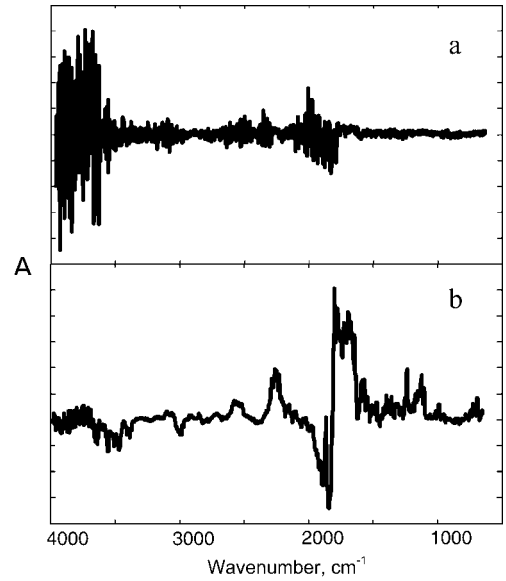

Figure 1. Sensitivity of carbonyl groups from (a) an unconstrained and (b) constrained ANN.

RESULTS AND DISCUSSION

All of the ANNs relied on one neuron in the output layer, with an output greater than 0.5 signifying the presence of a certain functional group; an output of less than 0.5 indicated the absence of this structure. Cross-validation was achieved by creating 10 different ANNs and dividing the total library into 10 subsets, each containing 275 IR spectra. Each network used a different subset as the validation set and trained with the remaining 90% of the library. Training ceased when the prediction error in the validation set was at a minimum. Because the weights of each ANN are initially random, some networks lead to better prediction results. In an attempt to compensate for this variation, 10 networks were created and tested for each of the validation sets, resulting in a total of 100 ANNs (10 networks with 10 subsets each). For each validation set, however, only the model that produced the lowest prediction error was retained for sensitivity analysis. The mean sensitivity was calculated from these 10 best networks.

Artificial Neural Networks Parameters. The first objective was to establish the importance of the constraints. A preliminary test was setup to predict the presence of the carbonyl group. Two kinds of ANNs were created, both with 10 nodes in the hidden layer. The first set of networks were unconstrained, whereas the second series employed the constraints of $t = 0.1$ and forced the weight vector to length of 1.0. The mean sensitivity for each of the ANNs is shown in Figure 1. Although the (Figure 1a) unconstrained network might have been more accurate with a prediction error of 0.58% as opposed to 1.02%, the unconstrained factor appears to be mostly noise. It is clear from the sensitivity

(28) Savitzky, A.; Golay, M. *Anal. Chem.* **1964**, *36* (8), 1627–1639.

Table 2. Prediction Errors Versus Number of Nodes in Hidden Layer for Carbonyls

no. of nodes	% error
1	1.71
5	1.34
10	1.02
15	1.26
20	0.94
25	0.98

factors that the (Figure 1b) constrained network's decision is based on the more relevant data, i.e., the carbonyl bands from 1830 to 1650 cm^{-1}. The negative band between ~1830 and 2000 cm^{-1} is due to summation bands in spectra of aromatic compounds. In spectra of aromatic compounds, a pattern of weak bands extends from about 1650 to 2000 cm^{-1}; thus, they overlap the carbonyl region. The negative contribution reduces the possibility that the weak bands in spectra of aromatics in the carbonyl region will be wrongly identified as carbonyls.

After determining the necessity of the constraints, the next goal was to vary the number of nodes in the hidden layer to find the model that produces the lowest prediction error. It follows that the most accurate constrained network should result in the most informative sensitivity factor. Again, a test was setup to predict the carbonyl functional group. The hidden layers contained 1, 5, 10, 15, 20, or 25 neurons. The errors (false positives plus false negatives) for each validation set using the different number of hidden neurons as well as the best results are listed in Table 2. The network with 20 hidden neurons was found to be most accurate with only 26 incorrect predictions, an error of 0.94%. As shown in Figure 2, the sensitivity factors with 10 or more nodes were quite similar, suggesting that the different networks are still learning the major trends associated with the carbonyl group.

The prediction accuracy for each functional group using the ANN analysis is shown in the second column of Table 3. The percent accuracy ranged from a minimum of 91.24 for C=C to over 99%. These prediction accuracies are higher than previously reported.[27]

Sensitivity Analysis. The ultimate goal of this research was to produce a spectral representation for each of the functional

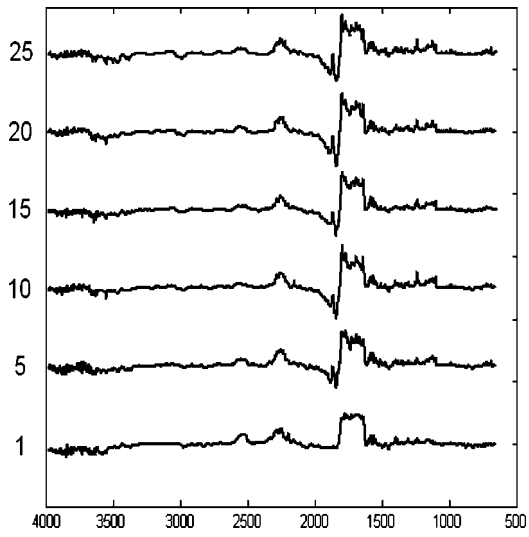

Figure 2. Carbonyl sensitivity using different numbers of hidden neurons.

Table 3. Prediction Errors and Accuracy Testing Spectral Library for Each Functional Group

functional groups	ANN % accuracy	sensitivity % accuracy
aromatic	92.62	91.61
amine	93.13	92.66
carbonyl	99.24	98.80
C−O	92.70	90.19
O−H	96.44	95.20
ketone	94.59	94.37
ester	97.38	96.73
aldehyde	97.82	97.13
acid	98.33	98.36
C=C	91.24	90.12
nitro	99.06	98.84
C≡N	97.06	96.40
alcohol	97.17	96.22

groups that could be used in a metric to predict the presence or absence of a functional group in future determinations. To facilitate this goal, a sensitivity factor as discussed in the theory section was produced for each of the functional groups in this study.

The sensitivity factor for the C=C group, as in alkenes, is shown in Figure 3. Although the highest prediction error was observed for alkenes, the sensitivity factor portrays its characteristic strong absorption bands as shown in Figure 3. The strong peaks of interest were due to the double-bond stretching at 1650 cm^{-1} and the four bands between 1000 and 850 cm^{-1} due to out-of-plane =C−H wagging motion.[29]

The sensitivity factors for the nitro and carbon triple bond nitrogen groups are shown in Figure 4. The prediction accuracy of the nitro group was over 99%. Although the nitro group was the least represented group in the entire study, the ANNs were able to emphasize the strong NO_2 stretching vibrations at 1550 cm^{-1} and from 1320 to 1380 cm^{-1}. The structure with the second fewest samples was the C triple bond N group. Figure 4b displays the sensitivity factor for this group and it is almost solely based on the nitrile C triple bond N stretch at ~2240 cm^{-1} and the isocyanide N triple bond C stretch at ~2140 cm^{-1}.

The sensitivity factors for (Figure 5a) ketones and (Figure 5b) aldehydes shown in Figure 5 are quite interesting, The sensitivity factors for both chemical groups have positive bands in the 1700 cm^{-1} region. However, aldehydes have a strong, positive band at 2725 cm^{-1}, and a weaker positive band at ~2810 cm^{-1}, whereas the ketone has similar negative bands at the same wavenumbers. The appearance of a doublet in spectra of aldehydes in the region of 2675−2850 cm^{-1} is due to Fermi resonance between the stretching vibration of the lone CH in aldehydes and overtone of the in-plane rocking vibration of the same CH bond, which appears

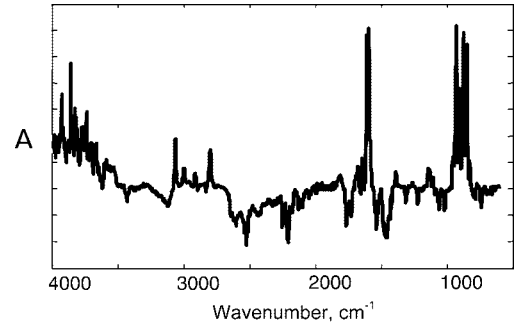

Figure 3. Sensitivity factor for alkenes.

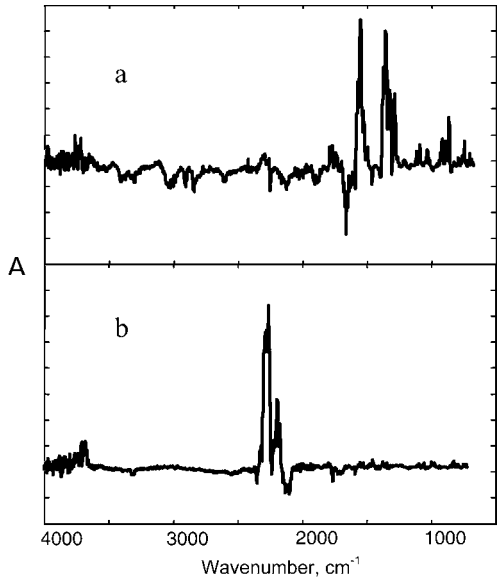

Figure 4. Sensitivity factors for (a) nitro and (b) C triple bond N functional groups.

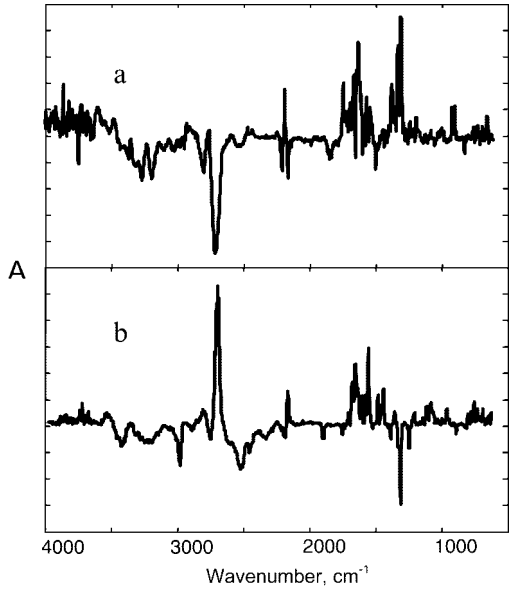

Figure 5. Sensitivity factor for (a) ketone and (b) aldehyde.

at ~1400 cm^{-1}. Jegla[22] observed a doublet of about equal intensity in the sensitivity spectrum of aldehydes when processing gas-phase spectra. However, the present study is on condensed phase spectra., and the broadness of other C−H stretching vibrations such as the CH_2 will overlap the CH stretch band in condensed phase spectra reducing the intensity of the C−H stretching band in the sensitivity spectrum. The negative intensity in the ketone factor suggests that it uses this region to exclude aldehydes from being misclassified as ketones and visa versa. The aldehyde has a sharp, negative band at 1320 cm^{-1}, and ketones have a strong positive band at the same wavenumber, again supporting the presence or absence of each group.

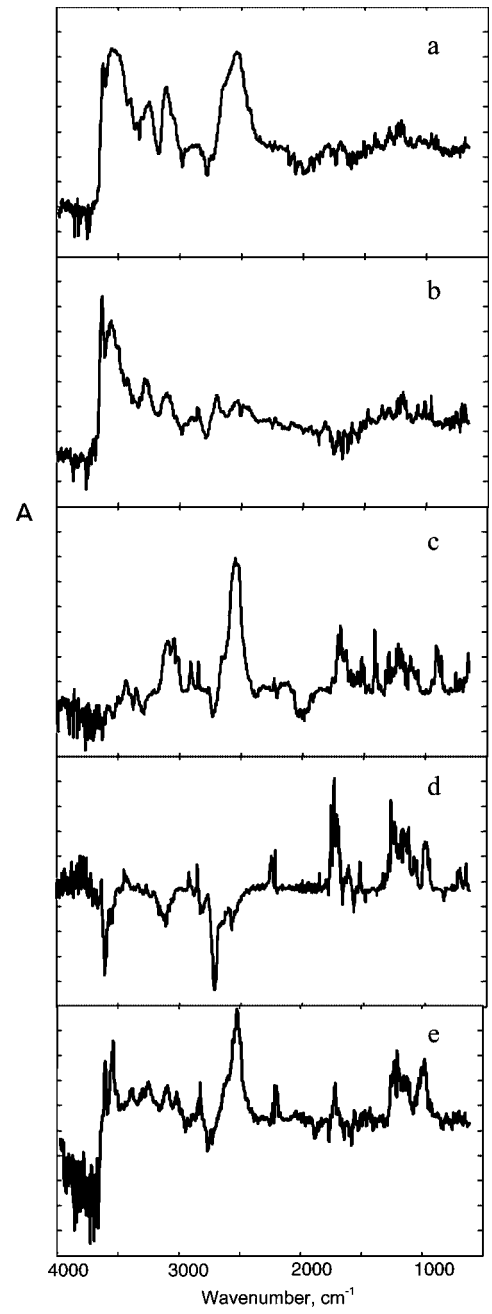

Figure 6. Sensitivity factors for (a) OH, (b) alcohol, (c) acid, (d) ester, and (e) −C−O−.

The sensitivity factors for OH, alcohols, acids, esters, and −C−O− are shown in Figure 6. The −OH sensitivity has the expected strong correlation between 2800 and 3800 cm^{-1}, but there is also a very sensitive band centered at 2525 cm^{-1}. The latter band was unexpected, but we find similar high sensitivities in the factor for acids and CO, which would suggest that it is due to an overtone of the fundamental C−O vibration in the 1200−1300 cm^{-1} region for acids. Upon closer examination of individual spectra, we find that organic acids generally have a weak absorption at ~2525 cm^{-1}. Esters (Figure 5d) have an intense negative band at 2690 cm^{-1}, whereas acids and −C−O− have a weaker negative band at about the same wavenumber. Acids, esters, and C−O have positive bands at ~1250 cm^{-1}, whereas esters and CO have a rather pronounced, positive band at 1000

(29) Lambert, J. B.; Shuvell, H. F.; Lightner, D. A. *Organic Structural Spectroscopy*; Prentice Hall, Inc.: Upper Saddle River, NJ, 1998.

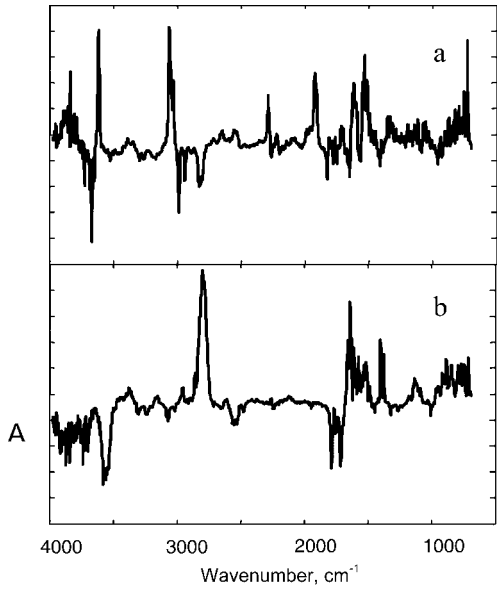

Figure 7. Sensitivity factors for (a) aromatic and (b) amine groups.

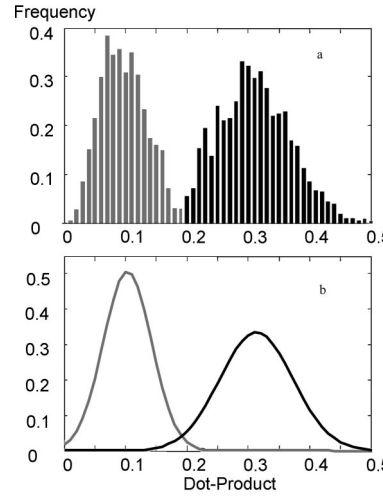

Figure 8. (a) Distributions for the (black) presence and (gray) absence of carbonyls in the IR library study and (b) the corresponding Gaussian fits.

Table 4. Average Confusion Matrix in Percentages for the IR Library

	IR	
known	*F*(pres)	*F*(abs)
present	19.5	2.6
absent	2.3	75.6

cm^{-1}. The presence of the −C−O− group in many different substructures might have been a reason for the relatively high prediction error in Table 3.

The aromatic factor shown in Figure 7a can be characterized by the C−H stretching bands just above 3000 cm^{-1} and the ring-stretching band between 1450 and 1630 cm^{-1}. The amine factor (Figure 7b) exhibits a strong band at ∼2800 cm^{-1} and an NH_2 deformation band around 1650 cm^{-1}. The very strong sensitivity band for amines at ∼2800 cm^{-1} is undoubtedly due to the so-called Bohlmann band. The existence of this band was first observed by Bohlmann[30] for cyclic immines. The band appears when α−CH bonds in heterocyclics have a dihedral angle of about 180° with the lone pair of electrons on the nitrogen atom.[31,32] The lower frequency and intensity of this band have been assigned to both lengthening of the C−H bond and to Fermi resonance with the overtone of the CH deformation.[33,34] The sensitivity spectra of both aromatics and amines exhibit sharp negative bands, which probably distinguish them from similar compounds. For example, the negative band in the sensitivity factor for amines at ∼3600 cm^{-1} distinguishes amines from alcohols. The negative bands in sensitivity factors can be extremely important for classifying different functional groups that have similar absorption features.

Decisions and Probabilities Using Sensitivity Factors. The final segment of this study was to use the sensitivity factors to predict the presence or absence of functional groups without the large, complex ANN. Two frequency of occurrence distributions were created by calculating the dot product between each library spectrum and the appropriate sensitivity factor. There will be one distribution for samples containing the functional group and another for those that do not contain the functional group. The resulting distributions for the carbonyl group are shown in Figure 8; there is a distribution of samples for each event, and either a carbonyl group is present or absent. The probability density functions usually have a Gaussian distribution;[37] thus, they were fitted with a Gaussian equation and normalized to an area of 1.0. These distributions represent the probabilities of obtaining a specific dot product given one of the instances, $P(D|h)$. Knowing the initial relative quantities of each functional group in the total library, $P(h)$, the conditional probability $P(h|D)$ can be determined using eq 6 and the final probability with eq 7.

Considering a normalized probability $P(h_{pres}|D)$ greater than 0.5 to indicate the presence of a functional group, the prediction accuracies from using the sensitivity factors are given in Table 3. As expected, using the sensitivity factors alone resulted in lower accuracy. In comparison to ANN, the difference is not that significant, with the greatest loss in accuracy found when testing the IR library for the C−O group, 92.70−90.19%. A confusion matrix consisting of the average results for the library is given in Table 4. The rows of this matrix represent the percentage of samples in which the functional group was present or absent; the columns represent the model's predicted results. For example, the average presence of a functional group in the library spectra was 22.1% (19.5% + 2.6%); the sensitivity/probability algorithm got 19.5/22.1 correctly identified as present and 2.6/22.1 wrongly identified as absent. The average absences of a functional group in the library spectra was 77.9% (2.3% + 75.6%); the sensitivity/

(30) Bohlmann, F. *Angew. Chem.* **1957**, *69*, 641; *Chem. Ber.* **1958**, *19*, 2157−2167.
(31) Mayo, D. W.; Miller, F. A.; Hannah, R. W. *Course Notes on the Interpretation of Infrared and Raman Spectra*; Wiley-VCH: New York, 2004.
(32) Krueger, J.; Jan, J. *Can. J. Chem.* **1970**, *48*, 3236–3248.
(33) Bertrand, B.; Nisole, C.; Drancourt, J.-M.; Dubuffet, T.; Bouchet, J.-P.; Volland, J.-P. *Spectrochim. Acta, Part A* **1996**, *52*, 1921–1923.
(34) Billes, F.; Geidel, E. *Spectrochim. Acta, Part A* **1997**, *53*, 2537–2551.
(35) McKean, D. C.; Duncan, J. L.; Batt, L. *Spectrochim. Acta, Part A* **1973**, *29*, 1037.
(36) McKean, D. C.; Ellis, I. A. *J. Mol. Struct.* **1975**, *29*, 81.
(37) Principe, J. C.; Euliano, N. R.; Lefebvre, W. C. *Neural and Adaptive System: Fundamentals Through Simulations*; John Wiley & Sons: New York, 2000; p 71.

Table 5. Average Probability for Correct and Wrong Predictions

	IR	
functional group	*P*(correct)	*P*(wrong)
aromatic	0.92	0.79
amine	0.93	0.80
carbonyl	0.99	0.85
C—O	0.93	0.78
O—H	0.97	0.82
ketone	0.94	0.76
ester	0.98	0.82
aldehyde	0.98	0.81
acid	0.99	0.82
C═C	0.92	0.75
nitro	0.99	0.81
C≡N	0.98	0.76
alcohol	0.97	0.77

probability algorithm got 75.6/77.9 correctly identified as absent and 2.3/77.9 wrongly identified as present. The false negatives are located in the top right, whereas the false positives are located in the bottom left. The errors were fairly evenly split between false negatives and false positives.

The average probability calculated for correct and incorrect classifications, both false positives and false negatives, are listed in Table 5. As evident in all cases, the average of the probabilities for misclassified samples was <0.85 (right column), whereas the average probability of correct predictions was >0.92 (left column). The significance of this is that in the event the model is correct it is confident in its decision, where in the event that the model is wrong it is less confident in its decision.

CONCLUSIONS

In this study, the fact that many functional groups consistently have bands in specific regions in infrared spectra was exploited. Artificial neural networks were employed to predict whether or not a chemical contained a certain molecular structure. We used sensitivity analysis on ANNs in order to extract relevant features. Here the partial derivative of the network's output with respect to each input variable was evaluated. The sensitivity factors illustrated regions of the spectrum that the network associated with each functional group, often correlating to real spectral features.

A simpler model for predicting functional groups was assembled using these sensitivity factors. The predictions of whether or not a sample contained a certain functional group were based on the dot products of each library spectrum with these sensitivity factors. The application of Bayes theorem added a probability to the model's decision. Instances in which the model made an incorrect prediction often correlated to a lower probability than when it was correct. This signifies the importance of providing a probability, a measure of confidence in the model's decision.

ACKNOWLEDGMENT

The authors are grateful to Peter R. Griffiths and John D. Jegla for making the latter's dissertation available prior to preparing the revised version of this manuscript.

Received for review January 8, 2008. Accepted March 25, 2008.

AC8000429