Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 2 Number 3 (2009) pp. 167–176

http://www.ripublication.com/jcib.htm

An Artificial Neural Network (ANN) to Identify Organic and Bio Molecules Based On Their Nuclear Magnetic Resonance (NMR) Spectra

Pablo Gomez and Claudia M. Lopez

Quantum Nanoelectronics, Inc., R&D Department, 12851 Luray Rd, Southwest Ranches, FL 33330, USA
E-mail: www.qnanotronics.com

Abstract

This paper reports a novel design and implementation of an Artificial Neural Network (ANN) aimed at recognizing unknown organic or biochemical compounds for which their Nuclear Magnetic Resonance (NMR) Spectra have been previously obtained. The purpose of developing this computational tool is to save time in chemistry and proteomics research. A chemist who needs to identify an unknown molecule for which its NMR spectrum has been recorded, can make use of this automated tool to retrieve the name and chemical structure of the closest known molecule matching the given NMR spectrum. This way, a researcher does not have to browse through large catalogs of NMR spectra to identify an unknown sample.

Introduction

NMR Spectroscopy is a widely used tool in chemical research. Each organic or bio molecule has a unique NMR spectrum that makes it possible to distinguish it from other compounds. There are numerous publicly available catalogs of NMR spectra of known compounds on the world wide web and other sources [1]-[4].

Based on these spectra the chemist is able to study the structure and properties of proteins, organic compounds and other molecules.

The ANN designed and implemented in this project is aimed at the automation of the process of recognition of an unknown sample using computer technology. Other similar studies do the opposite task, i.e., given a chemical structure or amino acid residue sequence, predict its NMR Chemical Shifts (Meiler, 2002), (Lisboa, 1998), (Baxevanis, 2005) whereas other ANN are specific to some chemical compounds such as alkanes (Ivanciuca, 1997) and olisaccharides (Svozi, 1995), (Studer-Imwinkelried,

2006). The ANN developed in this research was designed to be a general purpose tool for any kind of unknown chemical compound for which its ^{1}H spectrum is obtained.

NMR Spectroscopy Review

This section is intended for readers with a background in computer science or bioinformatics but with no prior knowledge of NMR. NMR Spectroscopy relies on the nuclear magnetic resonance phenomenon to study chemical compounds. The substance under investigation is first placed under a uniform magnetic field. The magnetic moments of nuclei of spin=1/2 align either in the direction of the applied magnetic field or in the opposite direction. This is due to the fact that there are only two valid quantum energy levels. The population of nuclei in the upper energy level, N_β, compared to the population of nuclei in the lower energy level, N_α, is the ratio (Cohan, 2005):

$$\frac{N_\beta}{N_\alpha} = e^{-\frac{\Delta E}{k_B T}} \tag{1}$$

where k_B is the Bolztmann constant and T is the absolute temperature in K. The resonance condition is satisfied when

$$h\nu = \Delta E \tag{2}$$

where h is the Planck's constant and v is the frequency of the electromagnetic radiation necessary for a transition from the lower to the upper energy level. This energy gap can also be represented in a more convenient way as (Friebolin, 2005)

$$\nu = \frac{\gamma}{2\pi} B_0 \tag{3}$$

where B_0 is the applied magnetic field and γ is the gyromagnetic ratio. This ratio is a constant for each type of nucleus. In NMR spectroscopy the two most common nuclei used are ^1H and ^{13}C for which the gyromagnetic ratios are 42.58 and 10.71 MHz per Tesla respectively.

The resonance condition is obtained by applying a controlled RF signal of a specific frequency.

However, nuclei in molecules are always surrounded by electrons and other atoms. The result is that in diamagnetic molecules the effective magnetic field B_{eff} is always less than the applied field B0 , that is, the nuclei are shielded. The effect, although small, is measurable and expressed by the equation

$$B_{eff} = B_0 - \sigma B_0 = (1-\sigma) B_0 \tag{4}$$

Here σ is a shielding constant. The resonance condition in equation (2) becomes now,

$$\nu = \frac{\gamma}{2\pi}(1-\sigma) B_0 \tag{5}$$

Figure 1 shows the 90 MHz ^1H NMR spectrum (B_0 =2.11T) of a mixture of bromoform ($CHBr_3$,3) , methylene bromide (CH_2Br_2,4), methyl bromide (CH_3Br,5)

and tetramethylsilane (TMS, Si(CH_3) $_4$,6)). The signal of TMS appears exactly at 90.00 MHz which is the reason why TMS is widely used as a reference chemical. Other compounds in the mixture have their resonant frequencies shifted with respect to the reference, and hence the name *Chemical Shift*.

In NMR spectroscopy a reference compound is always used, and the chemical shift is the resonant frequency difference with respect to this compound. This is expressed as

$$\delta_{sample} = \frac{\nu_{sample} - \nu_{reference}}{\nu_{reference}} \tag{6}$$

The numerator is typically no more than a few hundred Hz, whereas the denominator is usually several MHz. Therefore, the above definition yields very small numbers. For this reason it is customary to use parts per million (*ppm*), which is expressed as

$$\delta_{sample}[ppm] = \frac{\Delta\nu}{\nu_{reference}} \frac{[Hz]}{[MHz]} \tag{7}$$

In ^1H NMR Spectroscopy, the range of the chemical shift is typically less than 12 *ppm* whereas in ^{13}C NMR, this range is 200 *ppm*. The chemical shift is the most crucial parameter to design the ANN as will be explained in the next section.

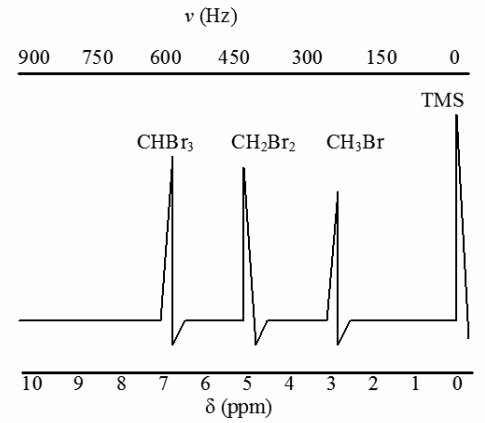

Figure 1: 90 MHz ^1H NMR spectrum of $CHBr_3$, $CH_2 Br_2$, $CH_3 Br$ and TMS.

Methods

A Backpropagation ANN was selected to classify and identify the NMR spectra. Backpropagation ANN is trained using supervision. What this means is that, for each training pattern presented to the input of the ANN, an associated target pattern is simultaneously presented at the output of the ANN.

Figure 2 shows the ANN model chosen to perform the task of identifying the NMR spectra. The ANN consists of 3 neural layers: input layer (vector X), hidden layer (vector Z) and output layer (vector Y).

Matrix V stores the trained weights of the hidden layer Z, while matrix W stores the weights of output layer Y after training.

The Backpropagation algorithm used to train the net uses random initialization of weights and the Nguyen-Widrow method to achieve faster learning (Fausett, 1994).

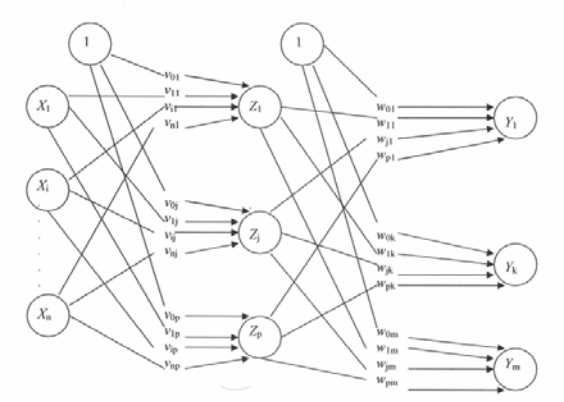

Figure 2: Backpropagation ANN with 3 layers.

The key parameters to design and tune the ANN are:

- The number of neurons in the input layer, N. This is based upon the nature of the pattern data that is going to be processed by the ANN.
- The number of neurons in the hidden layer, P. This number is empirically obtained after several training sessions. It is usually less than N.
- The number of neurons in the output layer, Y. This vector is also the result of the identification process. Therefore, the size and code of the output layer are such that uniquely represent all possible classes but the codes are sufficiently distinct to make the training faster and the net more accurate.
- The learning rate, alpha. This is a number between 0 and 1 that dictates how fast the weights are adjusted and how quickly the learning is achieved.
- The error threshold: this parameter indicates what is the maximum absolute error that is allowed. If it is too small, the training session will take longer but the effectiveness will be better. If it is too big, the training sessions are fast but the accuracy of the ANN is low.
- The number of epochs: this is the maximum number of cycles the training session will repeat the backpropagation algorithm on all training patterns. This is necessary when we need to stop the training session because the maximum absolute error never goes below the established error threshold.

Specific ANN Design Parameters

The most critical parameter of the ANN model is the quantization of the Chemical Shift (CS). The ^{1}H NMR Chemical Shift ranges from 0 to 12 ppm for most chemical compounds. The data used to design, train and test the ANN was provided by Dr. John Ralph with the University of Wisconsin at Madison. The resolution of the CS

data is 0.01 ppm, thus, the input vector to the ANN, X, was designed to have 12 ppm / 0.01 ppm or 1200 neurons or bins.

As an example, a small section of the ^{1}H NMR spectrum data of *Schistocerca gregaria chymotrypsin inhibitor complex* (BMRB database entry #6881) is shown in Table 1 and represented graphically on Figure 3. We chose a sequence of 18 amino acid residues of this polypeptide to explain how it is modeled and presented to the input of the neural network. The NMR spectrum provided as an example here corresponds to residues at sequences 7 through 24.

It is worth to mention that the same residue can have different resonant frequencies due to influences of neighbor residues. That is the case of GLY that appears 3 times at sequences 7, 20 and 23 with different values of ppm. Also, two different residues may have the same value of ppm as is the case of ASP at sequence 22 and LYS at sequence 24 with the same relative resonant frequency of 8.00 ppm.

Figure 3 shows the actual representation of the NMR spectrum used to feed the neural network. The data is stored on input vector X at their corresponding neurons, accumulating multiple occurrences of the same resonant frequency on the same neuron. As an example, at resonant frequency 8.00 there are two residues, thus, the number stored at its corresponding neuron number 800, is the value 2. Neurons having 1 resonant peak have a value of 1 and the rest of the neurons are set to zero.

Table 1: Sample ^{1}H NMR spectrum data.

Sequence	Residue	ppm	neuron number
7	GLY	7.28	728
8	THR	7.67	767
9	THR	8.55	855
10	PHE	8.98	898
11	LYS	8.77	877
12	ASP	8.19	819
13	LYS	9.38	938
14	CYS	8.81	881
15	ASN	8.32	832
16	THR	8.10	810
17	CYS	8.93	893
18	ARG	8.95	895
19	CYS	9.56	956
20	GLY	9.36	936
21	SER	8.89	889
22	ASP	8.00	800
23	GLY	7.79	779
24	LYS	8.00	800

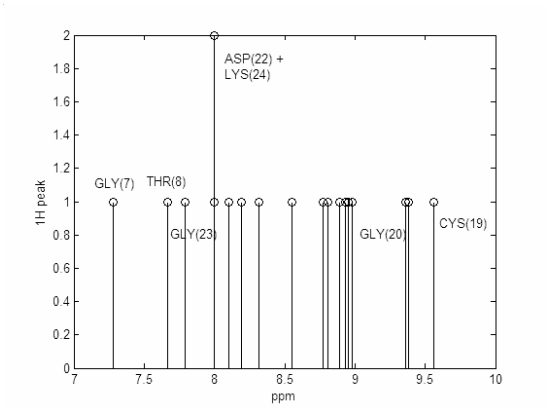

Figure 3: Sample NMR Spectrum.

Table 2: Quantization of Vector X.

Neuron number	Value
1	0
…	0
728	+1
…	0
767	+1
…	0
800	+2
…	+1/0
956	+1
…	0
1200	0

As a second example, let us model a hypothetical molecule that has the same residues as the one explained before but with 5 additional residues that are duplicates of the last 5 residues of the original one. The following tables and figure show the NMR Spectrum and quantized chemical shift stored on vector X.

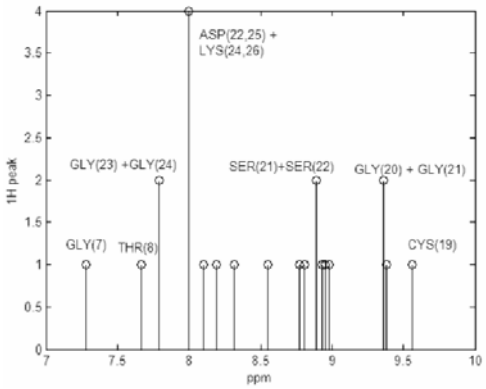

Figure 4: Hypothetical Molecule NMR Spectrum.

Table 3: Hypothetical Molecule ^{1}H NMR spectrum data sorted by ppm.

Sequence	Residue	ppm	neuron number
7	GLY	7.28	728
8	THR	7.67	767
23	GLY	7.79	779
24	GLY	7.79	779
22	ASP	8.00	800
24	LYS	8.00	800
25	ASP	8.00	800
26	LYS	8.00	800
16	THR	8.10	810
12	ASP	8.19	819
15	ASN	8.32	832
9	THR	8.55	855
11	LYS	8.77	877
14	CYS	8.81	881
21	SER	8.89	889
22	SER	8.89	889
17	CYS	8.93	893
18	ARG	8.95	895
10	PHE	8.98	898
20	GLY	9.36	936
21	GLY	9.36	936
13	LYS	9.38	938
19	CYS	9.56	956

Table 4: Quantization of Vector X for Hypothetical Molecule.

Neuron number	**Value**
1	-0
…	0
728	+1
…	0
767	+1
…	0
779	+2
…	0
800	+4
…	+1/0
936	+2
…	+1/0
956	+1
…	0
1200	0

The quantization of the CS translates a molecule's NMR into a data structure that uniquely represents it. No two molecules are likely to have the same representation. The molecule can have thousands of resonant peaks that are accumulated on their respective bins and thus provide an almost unique representation or NMR "fingerprint" of the compound.

The design of the output layer is a 32-bit bipolar (+1/-1) binary representation of each class of molecule. This way $2^{32}-1$ or 4,294,967,291 different molecules can be uniquely labeled. The unique binary value for each trained molecule is assigned randomly to avoid similar patterns. Using this arrangement an accuracy exceeding 95% was achieved.

One issue that had to be addressed was that to train an ANN to recognize an object, the ANN requires multiple patterns representing the same class of object. But in this case, only one single pattern is available to represent a given molecule, which is its NMR spectrum. This issue was resolved by programmatically creating 4 additional artificial patterns as shown in Table 5. These additional patterns are closely matched to the actual NMR spectrum and are also a way to simulate multiple slightly different spectra recorded for the same chemical compound. The maximum chemical shift variation of +/- 0.05 ppm corresponds to the margin of error reported by the data source used in the experiments. In this way, the neural network was trained to be flexible about frequency shift variations.

To verify the ANN design, 256 different known NMR spectra were used. After several trials, the optimal hidden layer size was 350.

Table 5: Additional artificial ^{1}H patterns.

Additional Pattern	CS (neuron number)
#1 CS shifted right 0.02 ppm	Original CS + 2
#2 CS shifted right 0.05 ppm	Original CS + 5
#3 CS shifted left 0.02 ppm	Original CS - 2
#4 CS shifted left 0.05 ppm	Original CS - 5

Final Design and Training Sessions

Table 6 shows the final ^{1}H ANN design parameters. The training process stopped at the maximum number of epochs, 20,000.

Table 6: ^{1}H ANN Design Parameters.

Parameter	Value
Input Layer Neurons	1200
Hidden Layer Neurons	350
Output Layer Neurons	32
Learning rate (α)	0.06
Threshold	0.02
M	0.7
Epochs	20,000

Results

The ANN was implemented using MATLAB®. The neural network was tested using manually entered unknown samples whose NMR spectra were slightly different to the ones used for training. The chemical shift was manually altered within the –0.05/+0.05 ppm range which is the margin of error of the original data.

For the first experiment, 26 unknown samples were presented to the ANN, and in all cases the ANN was able to correctly identify the right molecule, thus achieving an accuracy of 100%.

The experiment was repeated increasing the number of unknown molecules to 48. In this case, one of the molecules was not identified correctly, yielding an accuracy of 47/48 or 98%. The molecule that was not recognized had its spectrum distorted by +0.05 ppm and -0.05 for 10 of its residues. This situation is not very realistic and when the data was modified to have 4 of those residues within the -0.05 to +0.05 ppm range the molecule was correctly identified.

Several other experiments with different test sets with sizes between 20 and 80 unknown molecules yielded a minimum accuracy of 95%.

Acknowledgements

This work was partially supported by a grant from the U.S. Air Force Office of Scientific Research (AFOSR), FA9550-05-1-0232.

Conclusions

A Backpropagation Artificial Neural Network was designed, trained and tested successfully with excellent results. The neural network was trained and tested with real data from a Protein NMR bank. The model chosen for the ANN is simple yet powerful enough to achieve high accuracy in excess of 95%. It has been shown that this model can scale to a large catalogue of molecules by carefully quantizing the CS variable and adequately representing the target classes. The key components for this success are: an efficient quantization of the Chemical Shift for ^{1}H, a random binary representation code used at the output layer for the identification of the distinct classes of molecules and careful selection of the size of the hidden layer and training rate.

References

[1] BMRB BioMagResBank, Department of Chemistry, University of Wisconsin-Madison. http://www.bmrb.wisc.edu/

[2] SDBS - Spectral Database System**,** National Institute of Materials and Chemical Research, Japan, http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi

[3] Chemical Spectra and Spectra Data, University of Texas at Austin, http://www.lib.utexas.edu/chem/info/spectra.html

[4] Sally A. Ralph, John Ralph and Larry L. Landucci., NMR Database of Lignin and Cell Wall Model Compounds. http://ars.usda.gov/Services/docs.htm?docid=10491

[5] Baxevanis A. et al. (2005), Bioinformatics: A Practical Guide To The Analysis of Genes and Proteins, Wiley-Interscience, Hoboken, New Jersey

[6] Cowan B. (2005), *Nuclear Magnetic Resonance and Relaxation*, Cambridge University Press, New York.

[7] Fausett L. (1994), *Fundamentals of Neural Networks. Architecture, Algorithms and Applications.* Prentice Hall, New Jersey.

[8] Friebolin H. (2005), *Basic One- and Two-Dimensional NMR Spectroscopy*, Fourth Edition, Wiley-VCH, Heidelberg, Germany.

[9] Ivanciuca O., Rabine J. et al. (1997), 13C NMR chemical shift sum prediction for alkanes using neural networks, Computers & Chemistry, Vol. 1, Issue 6, p.p. 437-443

[10] Meiler J. et al. (2002), Using Neural Networks for ^{13}C Chemical Shift Prediction-Comparison with Traditional Methods, *Journal of Magnetic Resonance*, 157, 242-252

[11] Lisboa P.et al (1998), Assessment of Statistical and Neural Network Methods in NMR spectral classification and metabolite selection, *NMR in Biomedicine*, 11, 225-234

[12] Studer-Imwinkelried, M. (2006), NeuroCarb : Artificial Neural Networks for NMR structure elucidation of oligosaccharides, PhD Thesis, University of Basel, Faculty of Science

[13] Svozi D., Jirí Pospíchal et al. (1995), Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes, *Journal of Chemistry, Information and Computer Science*, 35, 924-928.