Artificial Neural Networks for the Classification of Meliaceae Extractives

Leigh-Anne Fraser

February 1998

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry, University of Natal, Durban, South Africa.

Abstract

The goal of this project was the development of a computer-based system using artificial intelligence to classify the limonoids, protolimonoids and triterpenoids isolated from the family Meliaceae by the Natural Products Research Group of the University of Natal, Durban. A database of samples was obtained between 1991 and 1996, part of which time the author was a member of the group and isolated compounds from *Turraea obtusifolia* and *Turraea floribunda*.

Over and above the problem of complexity and similarity in structures of the abovementioned natural products, are other difficulties. These include very small amounts of sample being isolated producing very weak peak signals in the C-13 NMR spectra, extraneous peaks in the NMR spectra due to different impurities and instrument noise, non-reproducible spectra due to the pulsed Fourier transform intervals and the nuclear Overhauser effect, impure samples often isolated as stereoisomeric mixtures or as mixed esters and superposition of peak signals in the NMR spectra due to carbons in the same environment within the same compound. These factors make identification by traditional computational and expert systems impossible. As a result of these shortcomings, the author has developed a novel approach using artificial neural network techniques.

The artificial neural network system developed used real data from the 300 MHz NMR spectrometer in the Department of Chemistry, Durban. The system was trained to discriminate between limonoids, triterpenoids and flavonoids/coumarins from the C-13 NMR spectra of pure, impure and unseen compounds with an accuracy of better than 90%. Further differentiation of the glabretals from the rest of the protolimonoids as well as from the rest of the triterpenoids showed similarly significant results. Finally, individual limonoid discrimination within the limonoid dataset was extremely successful.

Apart from its application to the extractives from Meliaceae, the methodology and techniques developed by the author can be applied to other sets of extractives to provide a robust method for the spectral classification of pre-identified natural products.

Preface

The research work described in this thesis was performed in the Department of Chemistry at the University of Natal, Durban, under the supervision of Professor D.A. Mulholland.

These studies represent original work by the author and have not been submitted in any form to any other university. Where use has been made of the work of others, it has been duly acknowledged in the text.

Signed: LA Fraser

Leigh-Anne Fraser (nee Akerman)

I hereby certify that the above statement is correct.

Signed: DA Mulholland.

Professor D.A. Mulholland

Acknowledgements

Primarily, I would like to thank my supervisor, Professor D.A. Mulholland for her invaluable advice, experience and knowledge as leader of the Natural Product Research Group.

Secondly, I am indebted to Dr D.D. Fraser for his invaluable assistance and encouragement in the field of artificial neural networks.

I would also like to thank Mr D. Jagjivan for running the NMR spectra and Dr Boshoff for his help with high resolution mass spectra.

Finally, I would like to thank the Foundation for Research and Development and the University of Natal for their financial assistance.

Table of Contents

Chapter 1

Introduction, Problem Statement and Thesis Summary

Chapter 2

Introduction to Limonoids and Protolimonoids

Chapter 3

Structural Elucidation of Limonoids and Protolimonoids

Chapter 6

Neural Networks Experimental

Chapter 7

Neural Networks Results

Chapter 1

Introduction, Problem Statement and Thesis Summary

1.1 Natural Product Chemistry

Natural product chemistry is a branch of organic chemistry that deals with the isolation of new compounds from nature, the so-called natural products. Originally, the unknown compounds could only be subjected to known chemical reactions in order to draw structural inferences. However, since the development of Nuclear Magnetic Resonance spectroscopy (NMR), extensive information about the structure of unknown natural products can be obtained in far less time. The common procedure for isolating natural products involves drying, grinding and extracting plant material with a suitably chosen solvent. Evaporation of the solvent then leaves a mixed gum or oil that must be separated into its component compounds. Gravitational column chromatography and flash chromatography (Naidoo, 1997) were used exclusively for the purification of the compounds isolated for this thesis.

The Natural Products Research Group at the University of Natal has devoted most of its research to the isolation of a particular type of natural product known as limonoids. Limonoids are fairly complex triterpenoid based compounds (see Chapter 2) which are frequently isolated in association with protolimonoids (limonoid precursors), other triterpenoids and flavonoids and coumarins. The structural elucidation of these compounds by means of NMR spectroscopy requires the interpretative ability of human experts who have gained their experience from years of research.

1.2 Problem Statement

There is a need to capture this human expertise at elucidation of these unique compounds. The author's intention was to develop a classification database of the ^{13}C NMR spectra of pre-identified limonoids as a tool to aid future inexperienced researchers, or as a basis for an NMR library as is commonly found with mass spectrometry instruments. This would thus enable the inexperienced researcher to confirm as to whether their isolated natural product was a limonoid, protolimonoid, glabretal-type protolimonoid, triterpenoid, flavonoid or coumarin. If a limonoid, the database would be able to confirm whether or not it had been isolated before and, if not, whether or not it had all the rings of the nucleus intact, α,β-unsaturation or a carbonyl group. At the outset, it must be stressed that this was a classification problem particularly applicable to natural products. It was neither an empirical elucidation, nor classification problem dealing with readily obtainable, synthetically pure chemicals. Thus, an intelligent computer-based classification technique was sought that was able to cope with the associated difficulties encountered in natural product isolation. These difficulties include:

1) obtaining the isolated compounds in very small quantities, resulting in very weak ^{13}C NMR spectra with peaks that could not be distinguished above noise levels
2) obtaining the isolated compounds in such small quantities that sufficient sample purification is not possible, resulting in extraneous "dirty" peaks in the ^{13}C NMR spectra
3) obtaining the isolated compounds in the form of inseparable stereoisomeric and mixed esters, resulting in doubling of certain peaks in the ^{13}C NMR spectra.

Furthermore, the number of resonances in the ^{13}C NMR spectrum does not always represent the number of carbon atoms in the compound due to peak superposition of carbon atoms in the same environment within the compound. Secondly, ^{13}C NMR spectra are not quantitative enough for classification by traditional computer classification techniques. Thus an absolutely precise, quantitative technique such as peak picking which involves correlating each carbon atom in the compound to a resonance in the ^{13}C NMR spectrum would not be an optimal solution.

Given the limitations and difficulties of natural product classification, as briefly discussed above, further complications were unique to our dataset of limonoids, protolimonoids, triterpenoids, flavonoids and coumarins in that, except for the flavonoids and coumarins, these compounds are all very large and complex (30-40 carbon atoms).

A more generalised technique that could cope with inconsistencies and irregularities was necessary. Artificial intelligence, which mimics human intelligence, is naturally more appropriate as it has robustness to cope with inconsistencies and the ability to generalise. A branch of artificial intelligence known as neural networks was attempted with very positive results. Artificial neural networks are adept at classification and pattern matching of spectral signals and thus can be deployed as classifiers of limonoids, protolimonoids, triterpenoids, flavonoids and coumarins. Pattern matching or pattern recognition is defined as the ability to assign an object to one of several possible categories, according to the values of some measured parameters. (Adams, 1995)

Initial research began when the author attempted to classify natural products, in particular limonoids and protolimonoids (isolated by the author from the species *Turraea obtusifolia* and *Turraea floribunda* which belong to the Meliaceae family) using artificial neural network computing techniques. Owing to the promising results of the initial research, the database of natural products was extended to include all those protolimonoids and limonoids isolated from the Meliaceae family by fellow researchers of the Natural Product Research Group at the University of Natal, Durban.

1.3 Thesis summary

Protolimonoids and limonoids are a particularly relevant natural product for study by the Natural Product Research Group, since they occur widely in the Meliaceae family, species of which are indigenous to and thus more accessible to a local group in Southern Africa. Therefore, Chapter 2 introduces, defines and explains the groupings and biosynthetic routes of limonoids (protolimonoids being precursors to

the limonoids). It also reveals the importance of limonoids by discussing their many and widespread applications.

Chapter 3 introduces some of the various elucidation techniques employed by the research group, as well as some of the difficulties associated with the elucidation of complex structures such as limonoids. ^{13}C NMR spectroscopy is discussed in detail and justified as to why the input data used for classification were taken from these spectra rather than the ^{1}H NMR spectra. The nuclear Overhauser Effect and pulsed mode Fourier transform NMR spectroscopy, which limit the reproducibility of ^{13}C NMR spectra, is also briefly explained (Kemp, 1991). Finally in Chapter 3, the elucidation procedure of two new prieurianin type limonoids, namely nymania-1acetate and a rohitukin type limonoid that were recently isolated by the author, are discussed, as they had not previously been documented.

Chapter 4 deals with some of the traditional computational classification techniques available, discussing briefly both supervised (Bayes' theorem, k-nearest neighbour) and unsupervised (hierarchical component analysis, principal component analysis, k-means algorithm, fuzzy cluster analysis) learning techniques. Human classification and Expert system classification were also discussed and the advantages and disadvantages of all three were stated.

Artificial Neural Networks as a classification technique are introduced in Chapter 5. Their computer simulation and development methods were detailed as well as their advantages and disadvantages. The application of neural networks in chemistry is described, as well as the suitability of neural networks to the classification of protolimonoids and limonoids. The use of a neural network simulation tool, *Predict* is presented. *Predict*, the multi-layer feedforward neural network application, which was used in the development of each classifier, is described as well as supplementary neural network classification techniques such as Learning Vector Quantisation (LVQ) and back-propagation neural networks that were used in some instances.

Experimental details are given in Chapter 6. The binning technique which converted NMR spectral data into inputs to the neural network is discussed along with other data preparation techniques, and it is shown how preliminary neural

networks were attempted using different categories, before the final Hybrid Neural Network structure was decided upon. The initial neural network (NN1) discriminated between limonoids, triterpenoids (which included the protolimonoids, glabretal-type protolimonoids, plant sterols, dammaranes and masticadienoic acids) and "other"(flavonoids/coumarins) classes. The separated triterpenoid class could then be run through NN2(a) or NN2(b). NN2(a) discriminated between protolimonoids (which included the glabretal-type protolimonoids) and the rest of the triterpenoids (which included the plant sterols, dammaranes and masticadienoic acids). NN2(b) discriminated between the protolimonoids, the glabretal-type protolimonoids and the rest of the triterpenoids. The separated limonoid data set could then be run through NN3(a) or NN3(b). NN3(a) could identify each of the known pre-identified limonoids from the ^{13}C NMR spectral data of both pure and impure limonoids, while NN3(b) could classify the limonoid data set as belonging to one of three classes. The structure, general parameters, data transformations, variable selection, learning rule, transfer function and evaluation technique are tabulated for each network in *Predict*, and the number of input, hidden (or Kohonen) and output neurons are tabulated for the LVQ and back-propagation networks.

Chapter 7 discusses the results of the various networks and parameters chosen in *Predict*, as well as the sensitivity and contribution analyses that were performed on the networks. In relevant instances, the results of *Predict*, LVQ and back-propagation neural networks are compared.

Finally, Chapter 8 discusses how neural networks and the binning pre-processing technique coped with the many difficulties associated with natural product purity and concludes that this artificial intelligence technique is an excellent solution to the classification of natural products.

1.4 Publications

The author has presented work on the limonoids of *Turraea obtusifolia* and *Turraea floribunda* at the Frank Warren National Organic Chemistry Conference, Gordon's Bay, South Africa, 1992; The International Conference on Botanical Diversity, Cape Town, South Africa, 1993 as a result of which she was listed as a contributor to the book, "Botanical Diversity in Southern Africa", edited by B.J.Huntley, Pretoria 1994; and the Frank Warren National Organic Chemistry Conference, Bloemfontein, South Africa, 1995.

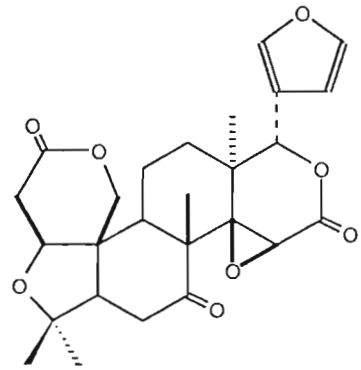
Journal publications have appeared in *Planta Medica*, Journal of Medicinal and Plant Research, 1992; *Phytochemistry*, 1994; the *South African Journal of Botany*, 1995 and *Phytochemical Analysis*, 1997.

Chapter 2

Introduction to Limonoids and Protolimonoids

2.1 Definition and Background to Limonoids

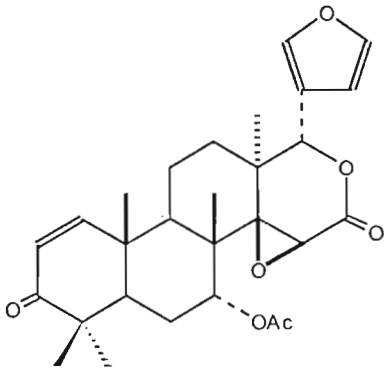
Limonoids appear to be derived from tetracyclic triterpenes that are widely distributed in the plant and animal kingdoms, and are defined as triterpene derivatives in which the side chain has become a furan ring by the loss of four carbon atoms. Hence they have an alternative name, the tetranortriterpenoids. Limonoids are named after the first limonoid ever discovered, called limonin (**1**).



LIMONIN (1)

The structure of limonin (**1**), an extractive of citrus fruits, was elucidated in 1960 by Arigoni *et al.* Citrus limonoids are responsible for the bitterness in the juice of citrus fruits and thus much work has focused on how to overcome this.

The first limonoid to be isolated in the Meliaceae family was gedunin (**2**). It was isolated in 1960 from *Entandrophragma angolense* (Taylor, 1984) which is a timber tree found in West Africa (Akisanya *et al.*, 1960). The structure was elucidated by comparison of its NMR spectra with those of limonin (Akisanya *et al.*, 1961).

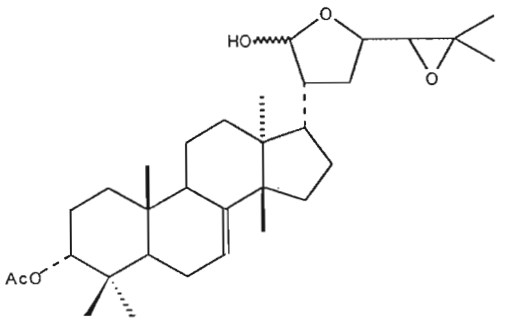


GEDUNIN (2)

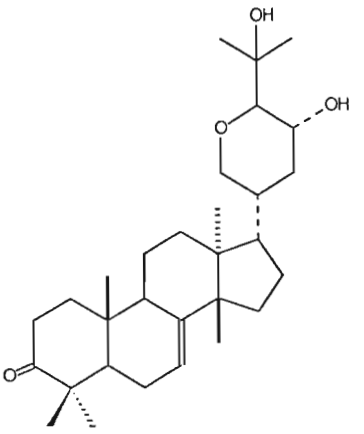
2.2 Protolimonoids

By their biological occurrence and oxidation pattern, the protolimonoids, which are also derived from tetracyclic triterpenes, appear to be biochemical precursors of the limonoids.

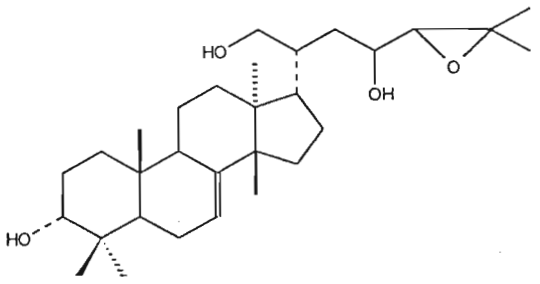
In the protolimonoids the eight-carbon triterpenoid side chain is intact but can be highly oxidised. In the case of turraeanthin (**3**), bourjotinolone A (**4**), entandrophragma triol (**5**) the sapelins (eg:sapelin B (**6**)), grandifoliolenone (**7**), and the glabretal-type protolimonoids (eg: glabretal, (**8**)), the side chain has cyclised to form an ether ring or a hemiacetal ring.



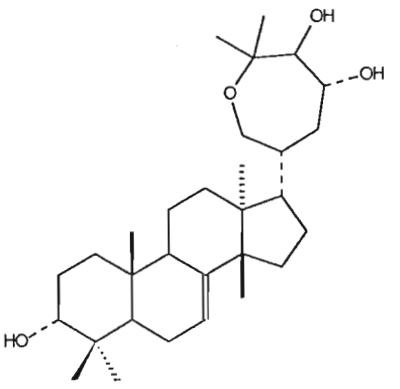
TURRAEANTHIN (3)



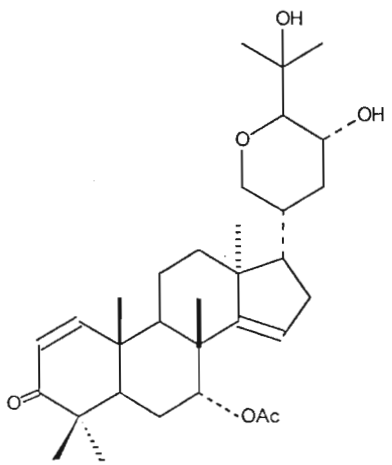
BOURJOTINOLONE A (4)



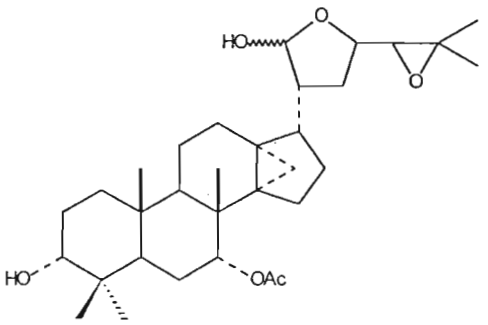
ENTANDROPHRAGMA TRIOL (5)



SAPELIN B (6)



GRANDIFOLIOLENONE (7)



GLABRETAL (8)

2.3 The Relationship between Protolimonoids and Limonoids

Limonoids appear to arise from the hypothetical triterpene precursor below (Connolly *et al.*, 1970); which is, in turn, oxidised to a protolimonoid, which undergoes further changes to give the simple true limonoids which are finally oxidised to the more complex ones as seen in Figure 2.1.

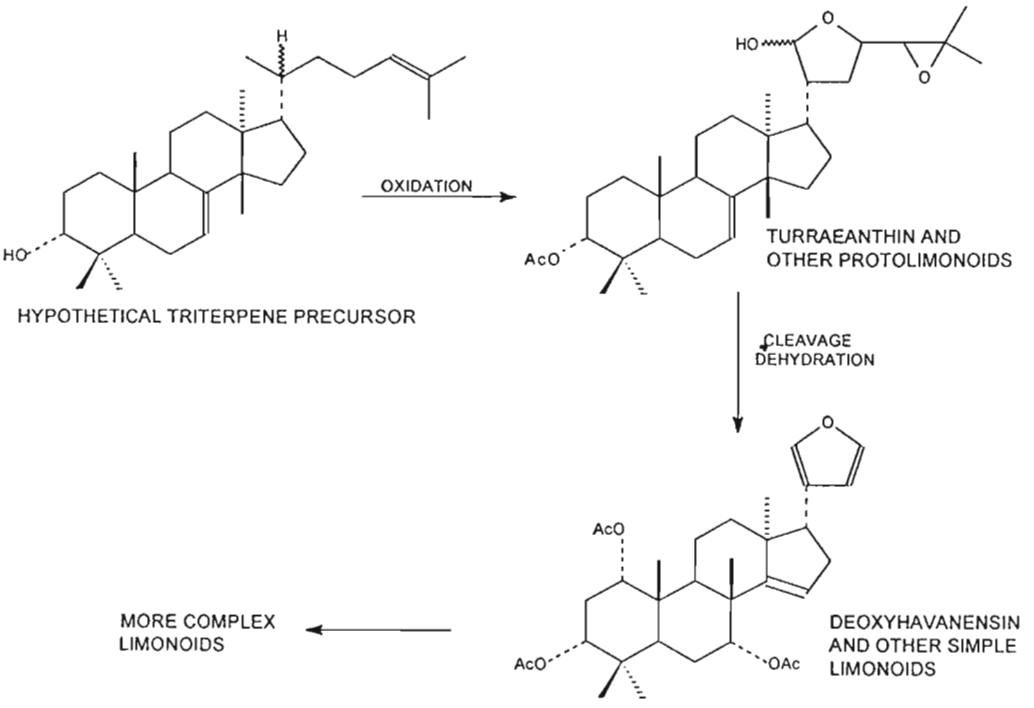


Figure 2.1 The biosynthesis of limonoids

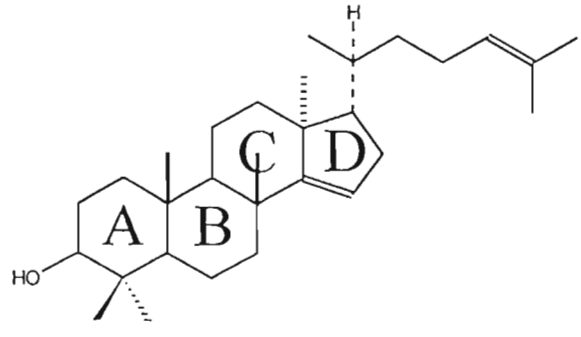
Ten main types of limonoids have been identified and they are represented in Table 2.1 as follows:

Table 2.1 The ten main types of limonoids

Group	Example	Ring A	Ring B	Ring C	Ring D	Side-chain
1	turraeanthin (3)	+	+	+	+	+
2	havanensin (20)	+	+	+	+	furan
3	khivorin (21)	+	+	+	lactone	furan
4	andirobin (22)	+	opened	+	lactone	furan
5	mexicanolide (23)	+	opened and recyclised	+	lactone	furan
6	phragmalin (24)	bridged	opened and recyclised	+	lactone	furan
7	evodulone (25)	lactone	+	+	+	furan
8	prieurianin (26)	lactone	opened	+	+	furan
9	nimbin (27)	+	+	lactone or opened	+	furan or further oxidised
10	obacunol (28)	lactone	+	+	lactone	furan

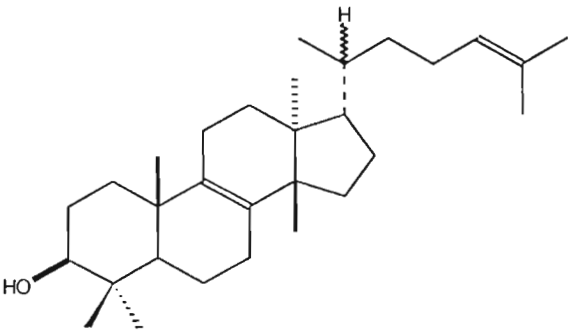
2.4 The Biosynthesis of Protolimonoids (including classification) and Limonoids

Arigoni *et al.*, (1960) proposed that limonoids were derived from an apo-euphol (**9**) structure, shown below.



APO-EUPHOL STRUCTURE (**9**)

The above structure was proposed to have been derived from either of the triterpenes euphol (**10**) or tirucallol (**11**).



EUPHOL (20 βH) (**10**)
TIRUCALLOL (20 αH) (**11**)

For the above to occur, two changes would be necessary:

2.4.1 The apo-euphol rearrangement (migration of the C-14 methyl group to C-8, and formation of a 14,15 double bond)

2.4.2 Cleavage of the C-23:C-24 bond and subsequent formation of a β-substituted furan ring

2.4.1 The apo-euphol rearrangement (migration of the C-14 methyl group to C-8, and formation of a 14,15 double bond)

The following scheme in Figure 2.2 outlines how Arigoni *et al.*, (1960) envisaged the formation of the apo-euphol (**9**) structure from tirucallol (**11**), euphol (**10**) or perhaps from butyrospermol (**12**), the Δ^7-isomer of euphol, by a similar rearrangement.

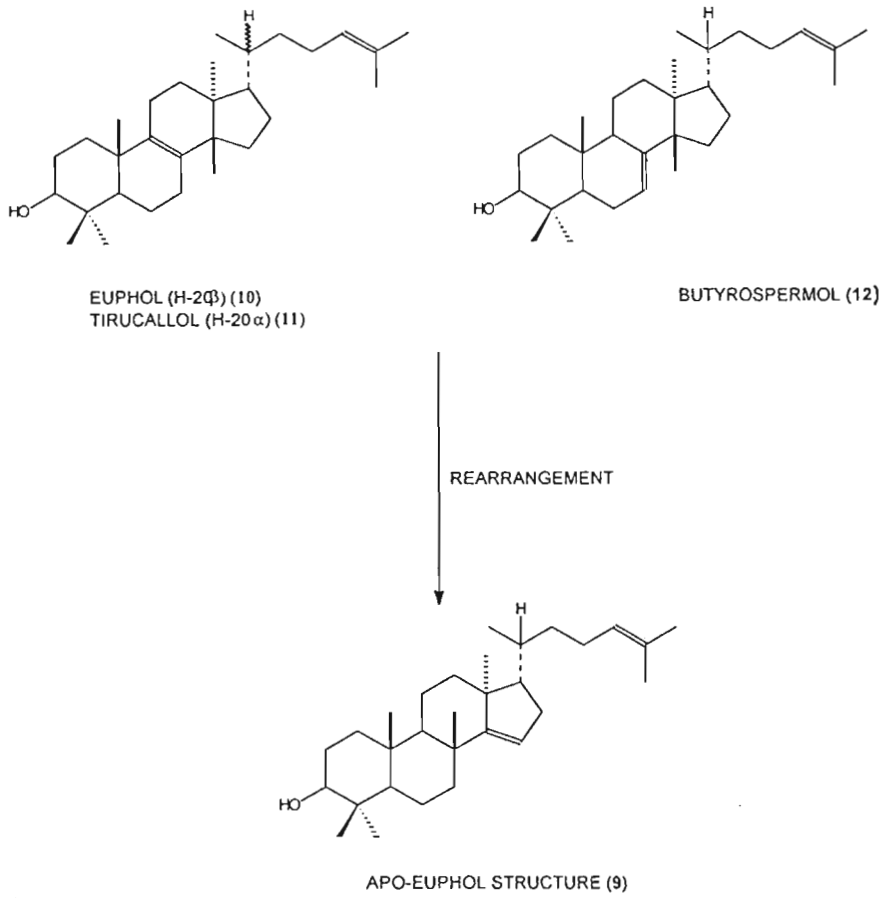


Figure 2.2. Formation of the apo-euphol structure

The known protolimonoids with the euphol (**10**) configuration occur in *Melia* with tirucallol (**11**) derivatives occurring in other genera (Taylor, 1984). An alternative formation of the apo-euphol (**9**) structure has been proposed directly from squalene (**13**) via the dammarene ion (**14**) as shown in Figure 2.3 (Moss, 1966; Cotterrell *et al.*, 1967):

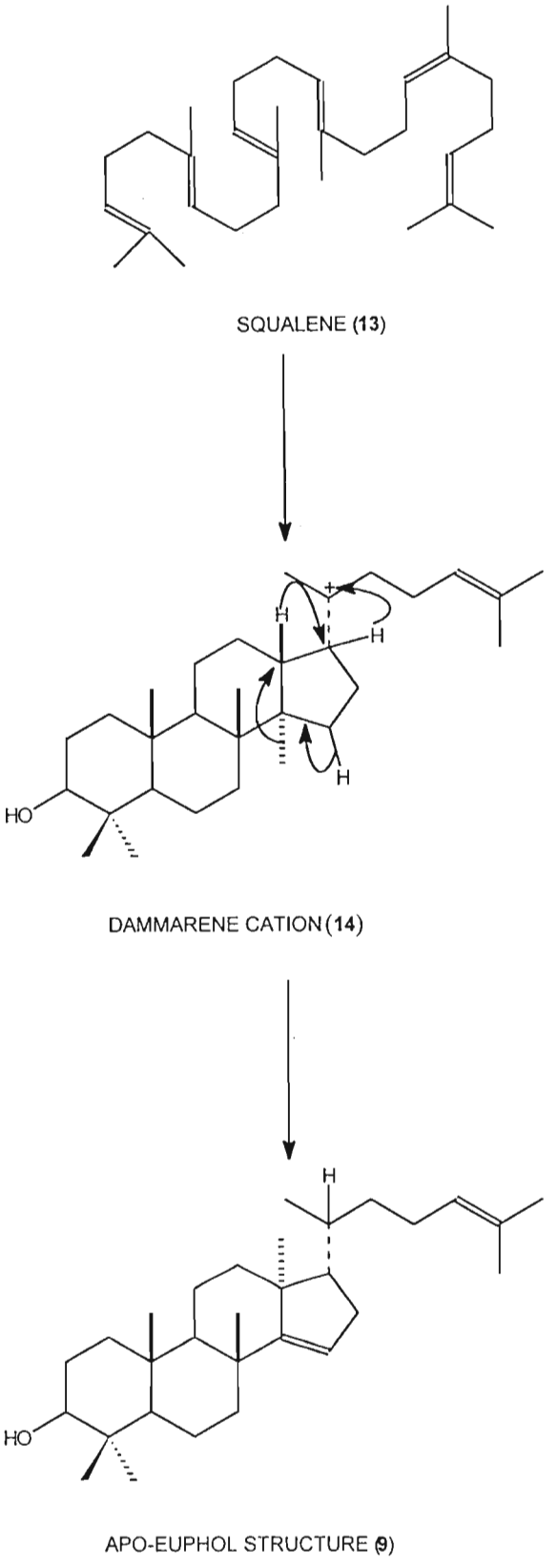


Figure 2.3 Formation of the apo-euphol structure via the dammarene ion

It is impossible to say whether limonoids are derived from the euphol (20β) (**10**) or tirucallol (20α) (**11**) precursor as C-20 becomes trigonal in the course of the rearrangement.

The isolation of turraeanthin (**3**) (Bevan *et al.*, 1967) which has the C-7:C-8 double bond and a side chain more oxidised than tirucallol (**11**), led to the suggestion that the limonoid precursor was tirucalla-7,24-dien-3β-ol (**15**). Cotterrell and co-workers, (1967) suggested a mechanism whereby this precursor could be transformed to the apo-euphol structure (**9**). All naturally occurring compounds having the apo-euphol (**9**) structure are oxygenated at C-7 and, when this takes the form of an hydroxy group, it is α-orientated. Thus a biogenesis involving the formation of the 7α,8α-epoxide of the hypothetical parent compound followed by the opening of the oxide ring with rearrangement to give the 7α-hydroxy apo-euphol (**16**) structure was proposed and shown in Figure 2.4:

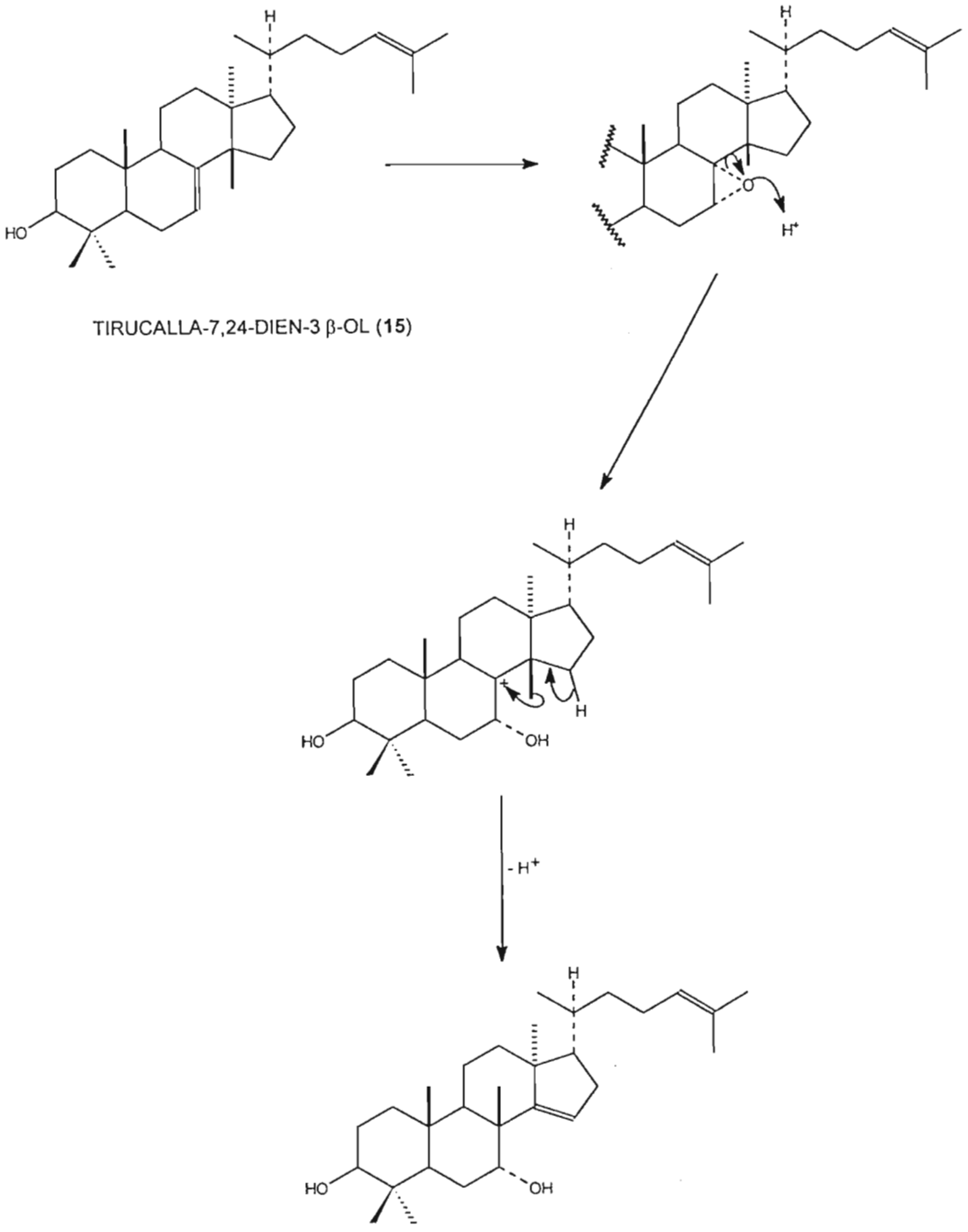


Figure 2.4 Formation of the apo-euphol structure via tirucalla-7,24-dien-3β-ol

Cotterrell *et al.*, (1967) demonstrated that the rearrangement of the 7α,8α-epoxide of methyl acetyl-Δ^7-dihydroelemolate (**17**) to give the 7α-hydroxy apo-euphol derivative (**16**) could be achieved in the laboratory using boron trifluoride-etherate as shown in Figure 2.5.

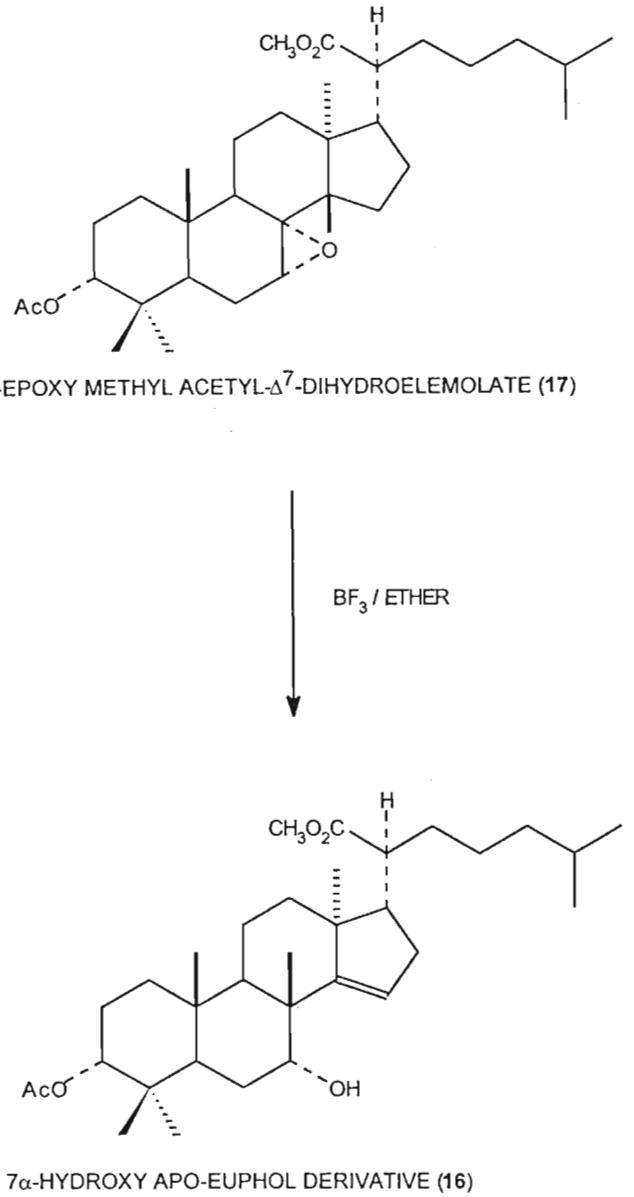
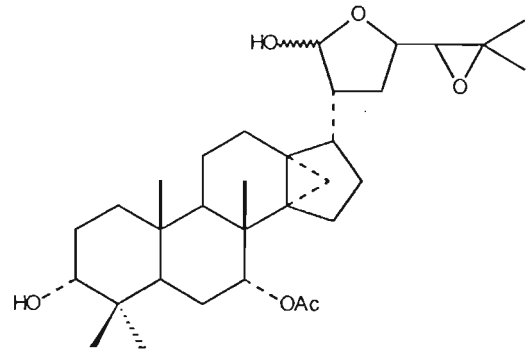


Figure 2.5 Laboratory synthesis of the apo-euphol structure

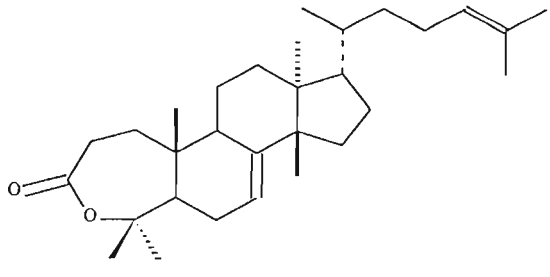
Thus protolimonoids fall in one of two classes. The first, like euphol (**10**), have a methyl group at C-14β and a double bond at Δ 7-8; while in the second, the so-called apo group, there is an hydroxy group at C7α, the double bond has moved to Δ 14-15, and the methyl group to C-8β.

Glabretal-type protolimonoids occupy an intermediate position between these two groups, as they have the 7α hydroxy and the 8β methyl group, but the 14,15 double bond is replaced by a 13,14 cyclopropane ring.

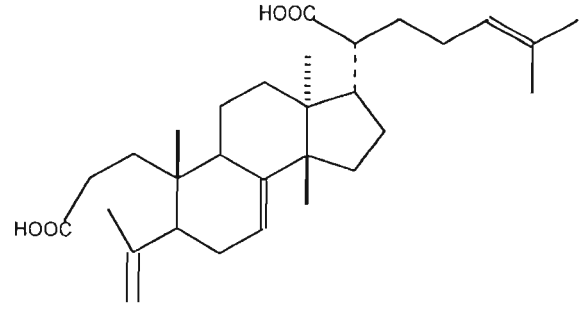


GLABRETAL (**8**)

In 1977 an interesting protolimonoid was isolated with ring A open (Okorie and Taylor, 1977). It was isolated from *Entandrophragma angolense* and given the name entandrolide (**18**). Later, compound **19** was isolated from *Entandrophragma delevoyi* (Roberts, 1994).



ENTANDROLIDE (**18**)



(**19**)

Thus protolimonoids can be grouped either according to whether the apo change has occurred or not or according to whether the side chain is oxidised or not.

2.4.2 Cleavage of the C-23:C-24 bond and subsequent formation of a β-substituted furan ring

The biochemical synthesis of the limonoids involves a series of oxidative changes, interspersed with molecular rearrangements. The oxidations are either epoxidations of double bonds or Baeyer-Villiger attacks on ketones. Although these oxidations are to be expected from a biological per-acid equivalent, laboratory duplication is sometimes difficult because of the extreme steric hindrance that the molecule is subject to *in vitro*. In the course of these changes, the triterpene side chain is first oxidised, eventually to a β-substituted furan ring by the loss of four carbon atoms. Hence the alternative name : tetranortriterpenoids.

Cotterrell *et al.*, (1967) proposed that the simple tirucallol side chain is oxidised in stages to produce an aldehyde group at C-21, a hydroxy group at C-23 and finally an epoxide in place of the C-24:C-25 double bond. The C-23 hydroxyl group cyclises onto the aldehyde carbonyl group to form the hemiacetal ring which is present in the turraeanthin side chain as shown in Figure 2.6:

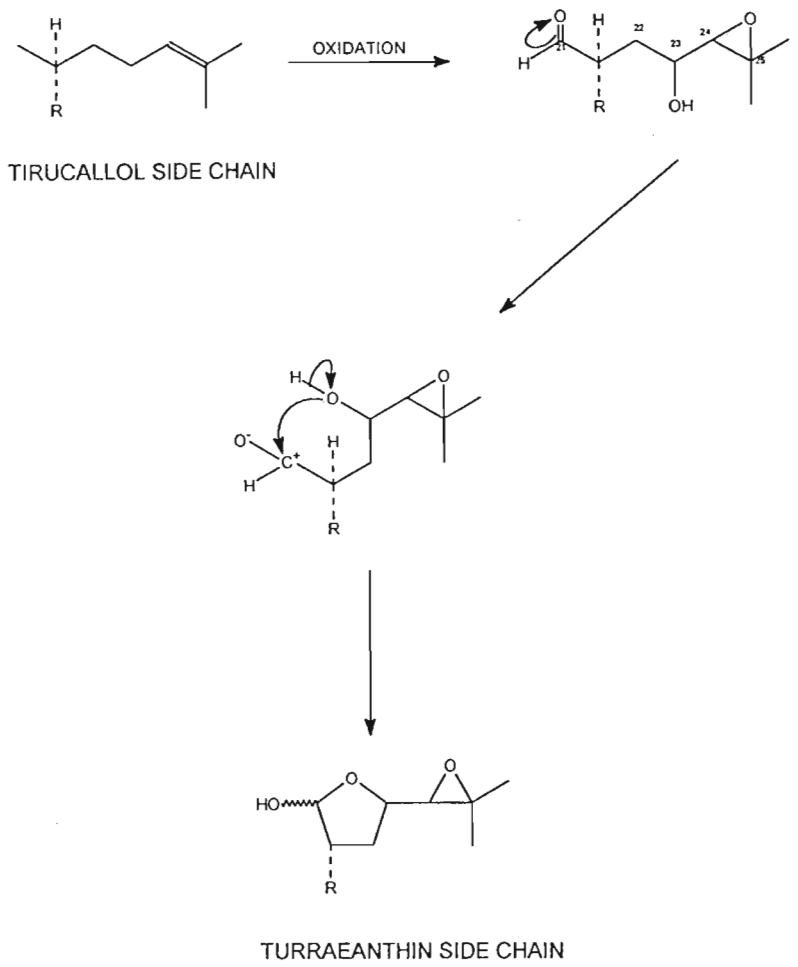


Figure 2.6 Formation of the turraeanthin side chain from the tirucallol side chain

Rearrangement of the epoxide could lead then to the formation of a keto group at C-24. This could also occur by the formation of a diol from the epoxide and subsequent oxidation of the C-24 hydroxy group (Cotterrell *et al.*, 1970). Fission of the C-23:C-24 bond of the side chain is believed to occur by a Baeyer-Villager oxidation producing the dihydrofuran which, on dehydration, would yield the β-substituted furan as shown in Figure 2.7:

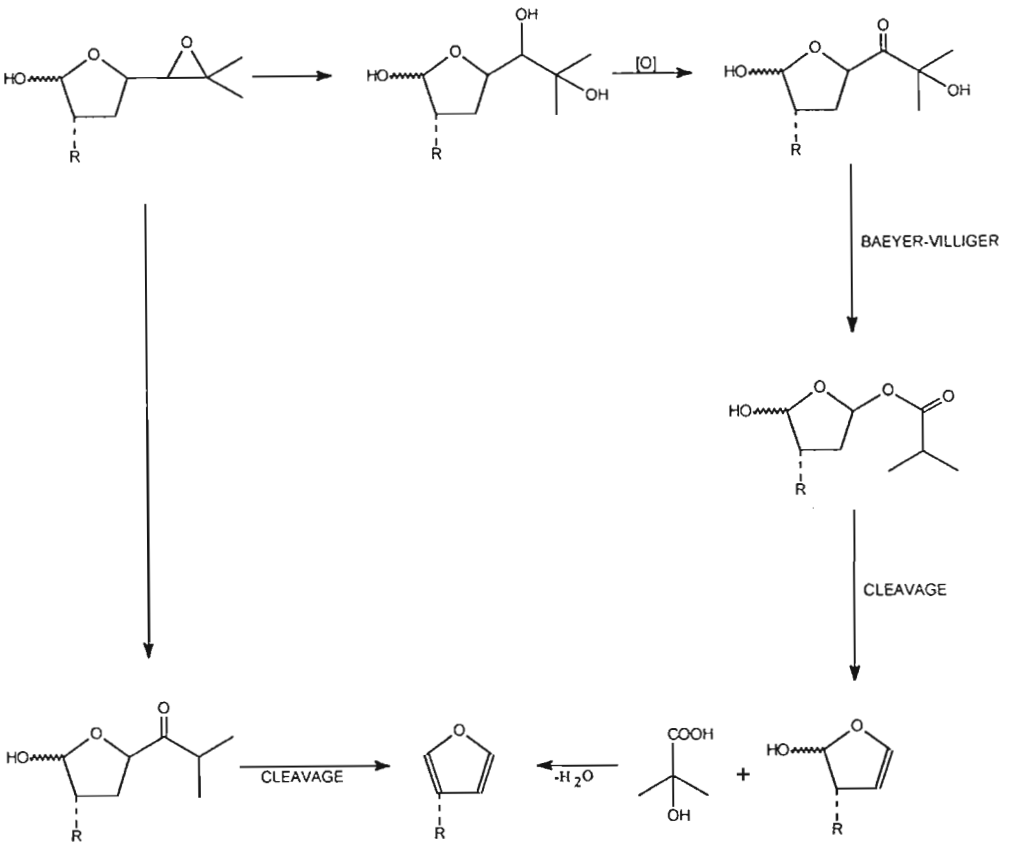


Figure 2.7 Formation of the furan ring from the turraeanthin side chain

The partial synthesis of the furan ring in havanensin-type compounds has been demonstrated in the laboratory by sodium metaperiodate oxidation of turraeanthin (3) in the presence of perchloric acid as shown in Figure 2.8, (Buchanan and Halsall, 1969):

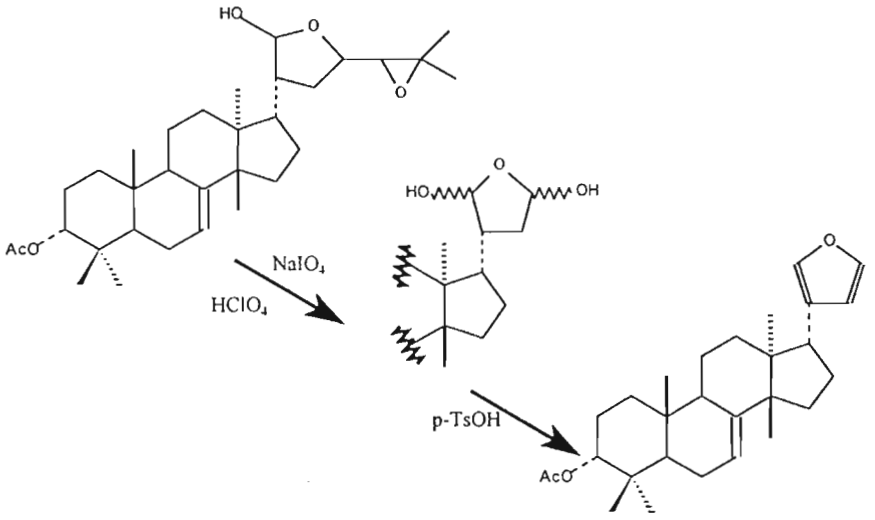
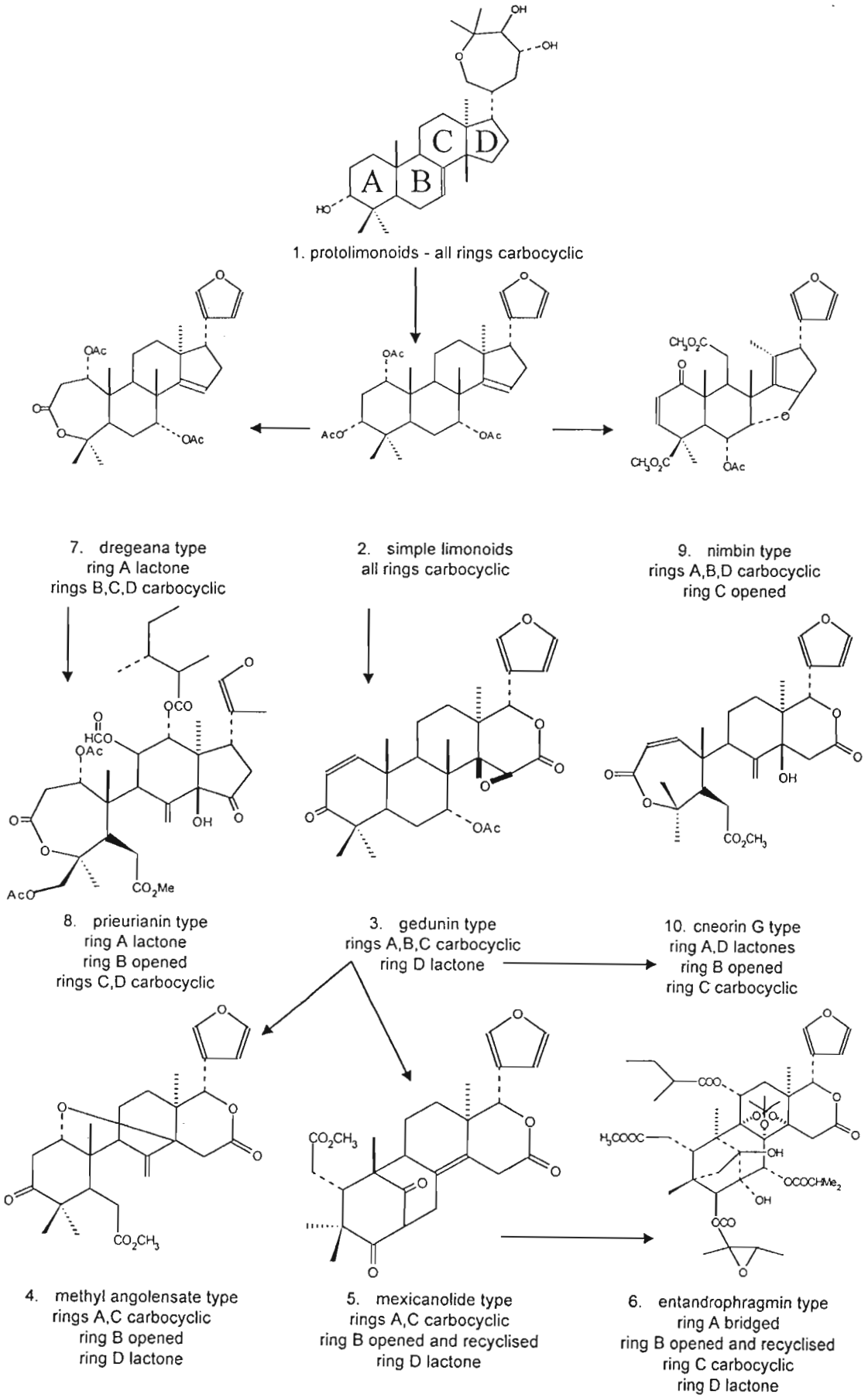


Figure 2.8 The partial synthesis of havanensin-type compounds

The biochemical relationships between protolimonoids and the different types of limonoids can be seen by consideration of the following Table 2.2 below:

Table 2.2 The biochemical relationships between protolimonoids and limonoids



Many limonoids are potentially available in very large quantities, the timber of some species may yield 1% of an isolated, crystalline limonoid while a single tree of *E. angolense* may contain more than 100kg of gedunin. A biosynthesis study of limonoids by tracer incorporation poses a problem as they are largely produced in the timber of large trees that is difficult to handle. Thus most research is done by extensive isolation of as many different limonoids as possible and by comparison of their structural and chemical properties. However, in the case of the citrus limonoids radioactively labelled precursor incorporation has been successfully utilised (Hasegawa *et al.*, 1986).

2.5 The Biological Importance of Limonoids and their Synthetic Uses

The biological reason for the production of limonoids in plants is most probably the fact that most limonoids are active as insect antifeedants, although not always directly insecticidal (Kraus *et al.*, 1978; Kubo and Klocke, 1981). Although it is true that insects would rather starve than eat leaves containing limonoids, the commercial applications may be limited due to the fact that insects may adapt to limonoids very quickly so that African insects eat African Meliaceae in preference to similar South American species and vice versa. Large scale trials to treat potato crops in Maine, USA have been conducted with limonin to protect them from the potato beetle that often destroys up to 25% of the crop (Mulholland, *pers comm*).

Limonoids have also been found to be active against cancer. For example, limonoids of the havanensin group, containing a 14,15 epoxide ring and members of the prieurianin group have been found to be active against some forms of cancer, although much of the work has been done on the Citrus limonoids, limonin and nomilin (Miller *et al.*, 1989; Lam and Hasegawa, 1987; Lam *et al.*, 1989).

Lam *et al.*, (1989) investigated the effect of Citrus limonoids on glutathione S-transferase (GST) activity. The GST enzymes are responsible for detoxification of xenobiotics, including carcinogens. Nomilin was found to be a potent GST inducer in the liver, small intestine mucosa and to a lesser degree in the forestomach. Limonin showed small inducing activity in the forestomach. As GST enzymes

serve to detoxify a system of carcinogens, it would be expected that there would be a correlation between enzyme inducing activity and inhibition of chemically induced tumour development. Lam *et al.*, (1989) investigated this with the use of benzo[a]pyrene (BP)-induced neoplasia in the forestomach of mice. Nomilin was found to be a potent inhibitor of BP-induced neoplasia in the forestomach of mice and limonin a weaker inhibitor.

Miller *et al.*, (1989) have investigated the effect of the limonoids limonin and nomilin on the development of 7,12- dimethylbenz[a]anthracene (DBMA)-induced buccal pouch epidermoid carcinomas in hamsters. The animals treated with limonin showed a 20% decrease in tumour number and a 50% decrease in tumour mass. The decreases with nomilin were less. The above findings suggest that some limonoids may be useful as chemopreventative agents.

Lastly, gedunin (**2**) the active constituent of *Cedrela odorata* (Meliaceae family) has been found to be active against malaria, the A-ring appearing to be the important factor (Khalid, 1989).

Limonoid research has proved to be quite significant in many areas of scientific advancement:

- They have proved to be active against cancer and malaria.
- They have also been found to be useful industrially as insect antifeedants.
- Isolation of the many varied types of protolimonoids and limonoids has contributed to a better understanding of the biochemical processes in plants.
- Their isolation has also been useful in the taxonomic classification of species within the Meliaceae family.
- Structural elucidation of the complex limonoids has contributed to the interpretation of NMR spectra of complex 3-D structures.

2.6 Summary

In this chapter, the close relationship between protolimonoids and the various classes of limonoids was revealed as well as their biosynthesis from simple triterpene precursors. Lastly, the importance of their uses was discussed so as to show their significance in natural product chemistry, and thus the value in setting up a classification database.

Chapter 3

Structural Elucidation of Limonoids and Protolimonoids

3.1 Introduction

In this work, structural elucidation implies the assigning of a chemical structure to a compound that has been extracted from some plant material. This is done by combining the results of various analytical chemical techniques, the most important of which is Nuclear Magnetic Resonance Spectroscopy (NMR). In order to classify these natural products, as is proposed in this thesis, competent structural elucidation must have taken place by experienced researchers.

3.2 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) involves the magnetic properties of certain atomic nuclei, notably the nucleus of the hydrogen atom (the proton) and that of the cabon-13 isotope of carbon (^{13}C). NMR spectroscopy is a form of absorption spectroscopy in which a sample can absorb radio-frequency electromagnetic radiation at frequencies governed by the characteristics of the sample. By studying a molecule by NMR spectroscopy, it is possible to record differences in the magnetic properties of the various magnetic nuclei present, and to deduce the position of these nuclei within the molecule. The number of different kinds of environments within the molecule can be deduced as well as which atoms are present in neighbouring groups. Usually, the number of atoms present in each different environment can be measured. A plot of the frequencies of the absorption peaks versus peak intensities constitutes an NMR spectrum.

3.2.1 ^{13}C NMR Spectroscopy

The nuclei that have magnetic properties will respond to the influence of an external magnetic field (B_0), and will tend to align themselves with the external magnetic field (the lower energy state) or against the external magnetic field (the higher energy state. The subsequent absorption of a beam of radio frequency results in the movement of the nuclei between the two states (Silverstein *et al.*, 1981). The ^{13}C nucleus has a spin quantum number, I of 1/2 whereas the commonly occurring ^{12}C nucleus is not magnetically "active"(spin number I, is zero). The number of orientations that the ^{13}C nucleus may assume is equal to $2I + 1$. Thus there are two energy levels with a slight excess of the lower energy state and it is possible to enter quanta of energy to effect transition between these two energy levels. The fundamental NMR equation correlating electromagnetic frequency with magnetic field strength is:

$$\nu = \tau B_0 / 2\pi$$

where,

ν = frequency [Hz]

B_0 = magnetic field strength [T]

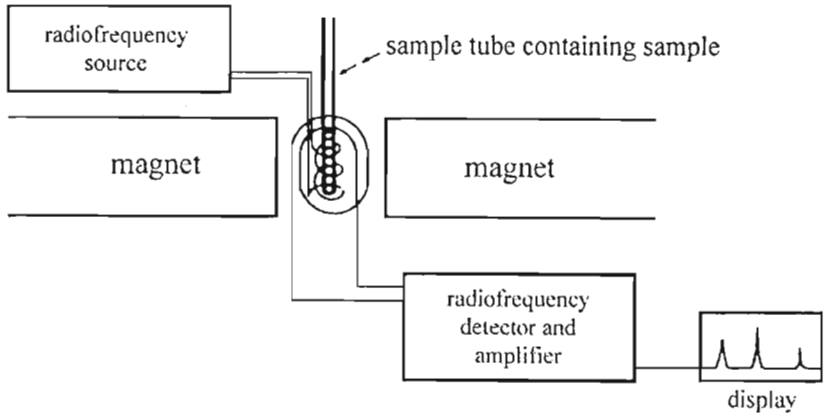
τ = magnetogyric ratio (fundamental nuclear constant)

Only a single peak should be obtainable from the interaction of radio frequency energy and a magnetic field on a nucleus in accordance with the basic NMR equation in which τ, the magnetogyric ratio, is an intrinsic property of the nucleus. Fortunately this is not so simple as each nucleus is shielded to a small extent by its electron cloud whose density varies with the electronic environment in the molecule. This variation gives rise to different absorption positions for each nucleus.

The magnetic moment of ^{13}C is one-quarter that of ^{1}H, so signals are inherently weaker, but the overwhelming problem in ^{13}C spectroscopy is that the natural abundance of ^{13}C is only 1.1% that of ^{12}C (Kemp, 1991). This has been overcome by the availability of pulsed Fourier transform instrumentation that permits simultaneous irradiation of all ^{13}C nuclei. The pulse duration may be of only a few microseconds, and when switched off, the nuclei undergo relaxation processes, and

re-emit the absorbed energies and coupling energies. As a result of all these simultaneously re-emitted energies, the instrument receives a complex interacting pattern that decays rapidly. This output is digitised in a computer and each individual frequency is identified from the interference pattern by the mathematics of Fourier Transforms and plotted on a linear scale. An entire spectrum can be recorded, computerised and transformed in a few seconds, and with a repetition every 2s, for example, 400 spectra can be accumulated in roughly 13 minutes achieving twenty times the signal enhancement.

NMR instrumentation basically consists of a strong magnet, a radio frequency transmitter, a radio frequency receiver, a recorder, calibrator and a sample holder that positions the sample relative to the main magnetic field, the transmitter coil and the receiver coil and which spins the sample thus increasing the apparent homogeneity of the magnetic field as shown in Figure 3.1. The ^{13}C NMR spectrum is presented as a series of peaks which all have varying frequency shifts from a reference marker.



(a)

Figure 3.1 The NMR Spectrometer (Kemp, 1991)

Limonoids, protolimonoids and triterpenoids are all organic molecules that consist of carbon, hydrogen and oxygen. Like carbon-13, hydrogen (also termed a proton) has a spin number I = 1/2 and a uniform spherical charge distribution.

Besides frequency shifts, coupling also occurs between neighbouring protons, and is denoted by a coupling constant, J, which is independent of the applied magnetic field. It is imperative that the chemical shift range far exceeds the coupling constants between adjacent protons. However, because of the large J values for ^{13}C-H (~ 110-320Hz), nondecoupled (proton coupled) ^{13}C spectra usually show complex overlapping multiplets that are difficult to interpret. Noise decoupling of the protons by means of double irradiation at their resonant frequencies by a wideband (noise) generator removes these couplings. This ensures that not only specific protons, but all protons are doubly irradiated simultaneously to form 1H-decoupled or noise-decoupled spectra. Thus, in the absence of other coupling nuclei, such as ^{31}P or ^{19}F, the result is a single sharp peak for each chemically nonequivalent ^{13}C atom, except for the infrequent coincidence of ^{13}C chemical shifts.

Decoupling can interfere with (shorten) relaxation times and, as a result there is an increase in signal (up to a factor of ~ 3) from the nuclear Overhauser effect (NOE). This enhancement results from the fact that the major relaxation route for a ^{13}C nucleus involves dipolar transfer of its excitation energy to the proton(s) directly attached to it. Thus maximum nuclear Overhauser effect operates on CH_3 followed by CH_2 and CH with no signal enhancement for quarternary carbons. This ensures easy identification of carbons.

Routine ^{13}C NMR Spectroscopy as was performed in this thesis is non-quantitative. There are two main reasons for this (Kemp, 1991):

1. The nuclear Overhauser effect tends to increase the line intensities for those carbons bearing protons, and to leave the quarternary carbons unaltered.

2. In the pulsed Fourier Transform mode used for normal ^{13}C work, the pulses are applied with only short delays between each successive pair. Carbon nuclei with long relaxation times will not have fully relaxed after one pulse before the next pulse is applied. The signals are therefore slightly saturated and of lower intensity. It is the quarternary carbons which tend to have long relaxation times, so that they show lowered intensities; in contrast to this, proton-bearing carbons

not only have shorter relaxation times, but also experience the enhanced line intensities caused by the nuclear Overhauser effect (Kemp, 1991).

A short, powerful radio frequency pulse excites all the ^{13}C nuclei simultaneously. Since the frequencies in the pulse are slightly off resonance (pulse centre at roughly the frequency of ^{13}C in TMS (tetramethylsilane) for all of the nuclei, each nucleus shows a free induction decay (FID), which is an exponentially decaying sine wave with a frequency equal to the difference between the applied frequency and the resonance frequency for that nucleus. Thus the FID display for a compound containing more than one ^{13}C nucleus consists of superimposed sine waves, each with its characteristic frequency and interference ("beat") pattern results. These data are digitised and stored, and a series of repetitive pulses, with signal acquisition and accumulation between pulses, builds up the signal. Fourier transform by the computer then converts this information to the conventional presentation of a ^{13}C NMR spectrum as shown in Figure 3.2.

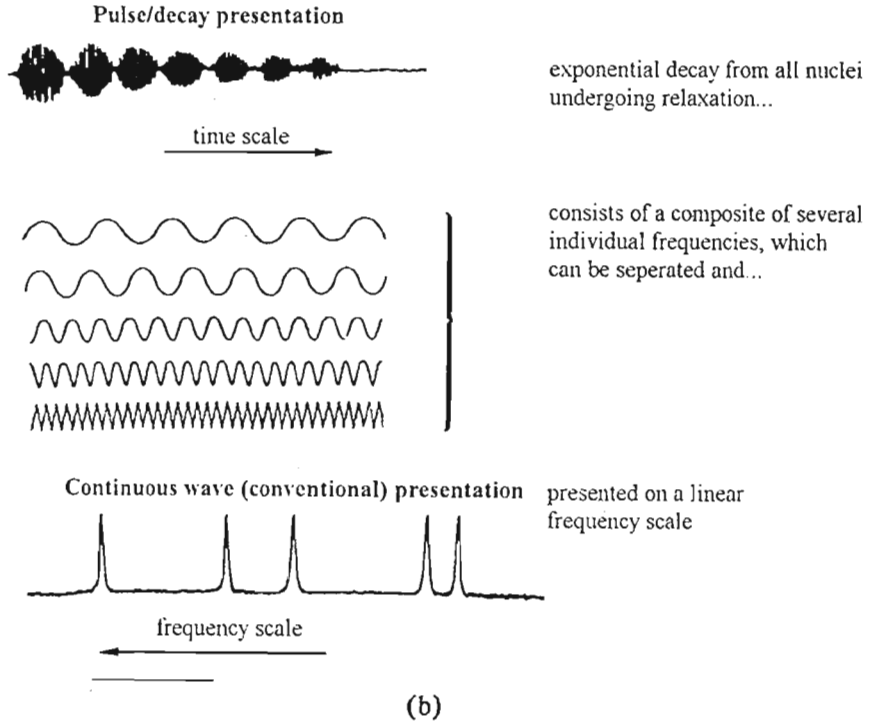


Figure 3.2 The process of obtaining the ^{13}C NMR spectrum (Kemp, 1991)

3.2.2 Other NMR Techniques

For total structure elucidation, however, a number of other NMR techniques must also be employed. These include:

PROTON SPECTRA

These spectra are based on the same principle of ^{13}C spectra but give more detail such as:

- the number of protons (hydrogens) in the compound
- the arrangement of CH_3, CH_2 and CH in the compound

DEPT SPECTRA (DISTORTIONLESS ENHANCEMENT BY POLARISATION TRANSFER)

These spectra make it possible to distinguish selectively the carbon-13 NMR resonances from CH_3, CH_2 and CH environments.

COSY (2D) SPECTRA (PROTON-PROTON CORRELATION SPECTROSCOPY)

The proton NMR spectrum is set out along the x-axis, and is repeated along the y-axis, with the signals repeated again in the contours of the diagonal peaks. Wherever a proton couples with another proton, i.e.: where correlation is established, it is indicated by the contour of an off-diagonal *cross-peak*. Thus the arrangement of hydrogens within the compound can be deduced.

HETCOR (2D) SPECTRA (HETERONUCLEAR CORRELATION SPECTROSCOPY)

This is a carbon-hydrogen chemical shift correlation spectrum in which the proton NMR spectrum appears on the y-axis and the carbon-13 NMR spectrum on the x-axis. Wherever correlation exists (that is, wherever a carbon and a proton are attached to each other), *cross-peaks* appear in the correlation spectrum. Thus it can be deduced which carbons are attached to which protons within the compound.

A range of other 2D NMR Spectroscopic techniques is currently used.

3.2.3 Other Elucidation Techniques Employed

Other confirmatory techniques include:

3.2.3.1 Mass Spectrometry

In simplistic terms, organic compounds are bombarded with electrons and converted to highly energetic positively charged ions called molecular ions, fragment ions and fragment radical ions which are separated in a variable magnetic field according to their mass and charge, and generate a current in proportion to their relative abundances. The mass spectrum is a plot of relative abundance against the ratio mass/charge (the m/z value) (Kemp, 1991).

These spectra convey the following:

- The relative molar masses of each compound with very high accuracy
- The places within a molecule at which it prefers to fragment so that the presence of recognisable groupings within the molecule can be deduced

3.2.3.2 Infrared Spectroscopy

When infrared light is passed through a sample of an organic compound, some of the frequencies are absorbed, while other frequencies are transmitted through the sample without being absorbed. The plot of absorbance or transmittance against frequency results in an infrared spectrum. Different infrared absorptions are associated with different functional groups, so infrared spectra are useful in identifying functional groups in an unknown compound.

3.2.3.3 Ultraviolet/Visible Spectroscopy

The absorption of ultraviolet/visible radiation by a molecule leads to transitions between the electronic energy levels of the molecule. The spectrum consists of a series of absorption bands, each band corresponding to an electronic transition. The strength of ultraviolet spectroscopy lies in its ability to measure the extent of multiple bond or aromatic conjugation within molecules. In general, differentiation can be made between conjugated dienes and nonconjugated dienes as well as α,β-unsaturated ketones from their β,γ-analogues.

3.3 Suitability of ^{13}C NMR Spectral Data for Classification

A study carried out by Maclachlan, (1981) on limonoid ^{13}C resonances showed that the resonance of a particular carbon atom depends largely on the local structure of the molecule (i.e. through a bond effect). Also, the position of that particular carbon resonance is fairly reproducible (within a range of about 8ppm maximum), in a variety of components. This fact makes the identification and structure elucidation of new compounds, by comparison of their carbon spectra with those already reported, a viable and extremely useful technique.

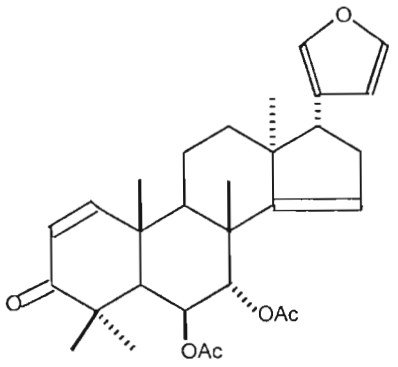
The same study indicated that the proton chemical shifts in limonoid 1-H spectra are more affected by the stereochemistry of the molecule, thus making the classification process more difficult.

Thus it was decided that if the whole spectrum was to be used as a 'signature' or 'fingerprint' for classification, use of ^{13}C proton-decoupled spectra should be used rather than the ^{1}H spectra as they are better resolved and the input vector can be more simply derived without the problems of peak multiplicity and peak areas.

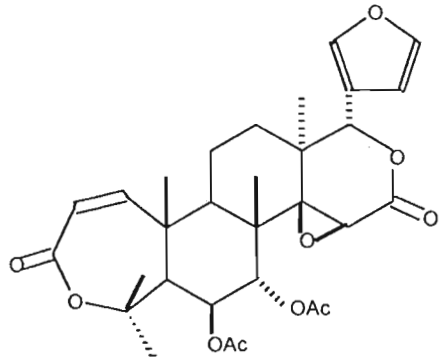
3.4 Intricacies of Limonoid Structural Elucidation

When isolated, limonoids may be of considerable molecular complexity, and the structure determination may be difficult. Several methods of determining stereochemistry have failed when applied to limonoids because of the complexity of their three-dimensional structures. Molecular conformation is not always readily predictable (Taylor, 1983), and use of isolated theoretical methods such as the Karplus equation and solvent shift methods has led to mistaken conclusions (Halsall and Trobe, 1975). The effect of substituents on nearby methyl groups (Jibodu *et al.*, 1970; Ohochuku and Taylor, 1969) has proved to be more reliable. The presence of a hydroxy substituent causes a considerable downfield shift of methyl substituents in a 1,3 diaxial relationship, while an acetoxy, or other carbonyl containing substituent, may or may not produce a similar shift depending on orientation effects. This produces good evidence of the stereochemistry, and in suitable cases of the position, of a substituent (Halsall and Trobe, 1975).

The incorrect structural assignment of dysobinin serves as an example (Singh *et al.*, 1978). Dysobinin was considered to be 6β-acetoxyazadirone (**29**). However, the lowest downfield methyl group resonates at δ1.3, while model compounds show that an axial acetoxy group at C-6 should produce a downfield shift of the methyl group at C-10 to approximately δ1.5. The 6α-configuration is predicted to lead to the observed shift (Connolly *et al.*, 1959). The structure of 6β-acetoxyobacunol (**30**), although unusual, appears to be correct as the methyl shifts recorded agree with those calculated.



6β-ACETOXYAZADIRONE (**29**)



6β-ACETOXYOBACUNOL (**30**)

3.5 Limonoids from Meliaceae

Limonoids have been found extensively within the Meliaceae family, as well as occurring in the Rutaceae and Cneoraceae families, although the Meliaceae differs from the latter two families in diversity of structures of its limonoids (Ansell, 1986).

Owing to its accessibility in the Southern African region, the Natural Products research team has focused its limonoid extractions on the very diverse family Meliaceae, of which the true Spanish Mahogany is a member. Previous research on the species *Turraea floribunda* which belongs to the Meliaceae family has yielded limonoids with structures related to the heudelottins and havanensin (Maclachlan, 1981) as well as two new limonoids (Fraser *et al.*, 1994) which were isolated from the seeds and which have ring B opened while not having the oxidatively opened ring D as those within the toonafolin group.

T. obtusifolia is closely related to *T. floribunda*, also belonging to the subfamily Melioideae and the tribe Turraeeae. *Turraea obtusifolia*, commonly known as the Small Honeysuckle Tree, is a small deciduous tree, only reaching approximately 3m in height, and although it is widespread, it is not common or abundant in any specific area, thus making it extremely difficult to obtain. To add to this problem, the tree does not bear prolifically. This species has been found to grow from coastal dunes to rocky hills.

Previous work on *T. obtusifolia* (Akerman, 1991) has yielded a number of protolimonoids and recent work on the seed by the author has yielded two limonoids, nymania-1 and a rohitukin type compound as well as a number of prieurianin type limonoids which were of insufficient quantity for comprehensive, rigorous elucidation.

3.5.1 Identification of nymania-1 and a rohitukin type limonoid from the seed of *T. obtusifolia*

The seed of *Turraea obtusifolia* was examined, and found to contain a rohitukin type limonoid as well as the limonoid nymania-1 (**I**), the principal limonoid found in *Nymania capensis*. This provides chemical evidence for a close relationship between the species *Nymania* and *Turraea* and supports *Nymania's* placement in the tribe Turraeeae of the Meliaceae family.

The isolation and elucidation process of nymania-1 (**I**) and the rohitukin type limonoid (**III**) is included in this chapter as an illustration of the structural elucidation of limonoids.

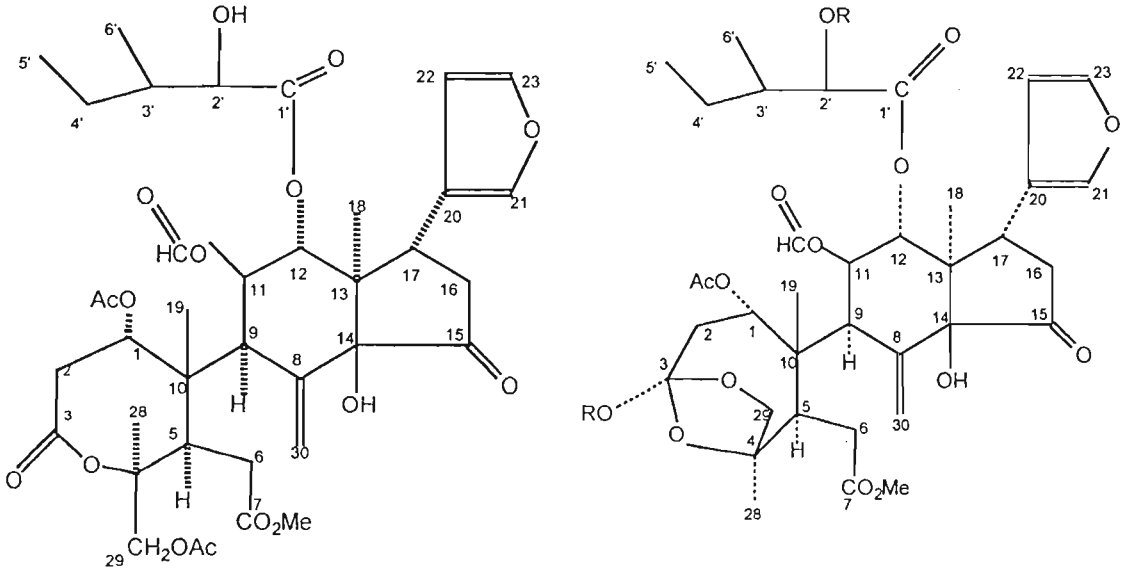
3.5.1.1 Materials and Method

Some of the seed material (18,97g) was collected along the South Coast of Natal while the other portion (108.45g) was obtained from Kirstenbosch Gardens. The seeds were dried, crushed and extracted separately in a soxhlet apparatus with refluxing hexane for 24 hours. The crude extracts were evaporated to dryness and the resulting residues (224,6mg and 5.82g respectively) were chromatographed on a gravity silica gel column (Merck. Art. 7726) using a solvent system of CH_2Cl_2 :

EtOAc :: 60 : 40 to yield pure nymania-1 (**I**) and rohitukin-type limonoid (**III**). Nymania-1 (**I**) was acetylated by dissolving it in 1,5cm^3 of pyridine, adding 1,5cm^3 of acetic anhydride and leaving it to react for 12 hours. Methanol was then added to terminate the reaction, after which toluene was added to remove traces of pyridine. The solvents were then removed using a rotary evaporator. Thin layer chromatography of the resulting mixture using CH_2Cl_2: EtOAc :: 60 : 40 showed, after spraying with anisaldehyde spray reagent, the major product at R_f = 0,9 along with some minor components. Repeated flash chromatography followed using a silica gel column in which the major component (13mg) was isolated. Structural elucidation of the compound took place using ^1H, ^{13}C, COSY and HETCOR nuclear magnetic resonance spectroscopy.

3.5.1.2 Results and Discussion

The structures of these limonoids were elucidated chiefly by comparison with prieurianin (**26**).



26

I R=H
II R=Ac

Figure 3.3 The three limonoids prieurianin (**26**), nymania-1 (**I**) and nymania-1 acetate (**II**).

3.5.2 Nymania-1 Acetate

The carbomethoxy group at δ 3.65 in the ^{1}H spectrum (ring B opened), the ring A lactone singlet at C-3 (δ 120.9) and ring D ketone at C-15 (δ 207.02) in the ^{13}C spectrum, the formate proton resonance at δ 7.95, three furan bands at δ 7.30, δ 7.30 and δ 6.20, and acetate and methyl group resonances indicated a prieurianin type limonoid. Compound **I** however, had only one acetate peak compared with the two acetate peaks of the prieurianin spectrum (**26**). Since compound **I** was amorphous and slightly impure, it was acetylated to yield the crystalline diacetate, $C_{40}H_{52}O_{17}$, (**II**) where acetylation had occurred on the tertiary hydroxyl at C-3 and in the C-12 side-chain ester. This was consistent with the results obtained by Maclachlan and Taylor, (1982) for the structure of nymania-1 (**I**). Comparison of ^{13}C NMR spectral data confirmed that the acetylated compound was nymania-1 acetate (**II**).

The three acetate methyl proton singlets were evident in the ^{1}H NMR spectrum at δ 2.05, δ 1.99 and δ 1.95, while the superimposed doublets at δ 5.90 (d, 11.5Hz, 1H), δ 5.98, δ 5.92 (2H) were assigned to H-12 and 2H-30, the multiplets at δ 5.27 (dd), δ 5.17 (dd) to H-11 and H-1, and the series of resonances at δ 4.27, δ 3.95, δ 4.08 and δ 3.70 to H-9, H-17, H-29A and H-29B respectively.

The coupling between H-9, H-11 and H-12 was evident from the COSY spectrum, H-9 (d, 7.7 Hz) being split by H-11β which appeared as a double doublet due to splitting by both H-9 and H-12. Weak coupling was observed between H-11β and the formate proton, hence confirming the latter's position. H-12α also appeared as a doublet (11.5 Hz) from splitting by H-11β.

2H-29 most often occurs as an AB quartet, in the region δ 3.5- δ 4.5 and the doublets at δ 4.08 (d, 7.7Hz, 1H) and δ3.78 (d, 8Hz, 1H) were consistent with this.

Table 3.1 ^{13}C NMR Data for nymania-1 acetate (**II**) $C_{40}H_{52}O_{17}$:

Carbon atom number	
C-1	71.11 (d)
C-2	39.12 (t)
C-3	120.90 (s)
C-4	85.44 (s)
C-5	49.44 (d)
C-6	34.06 (t)
C-7	170.14 (s)
C-8	138.28 (s)
C-9	50.60 (d)
C-10	49.09 (s)
C-11	72.12 (d)
C-12	74.10 (d)
C-13	49.44 (s)
C-14	81.10 (s)
C-15	207.02 (s)
C-16	41.68 (t)
C-17	35.26 (d)
C-20	123.05 (s)
C-21	140.99 (d)
C-22	110.59 (d)
C-23	142.98 (d)
C-29	73.76 (t)
C-30	126.28 (t)
C-1'	175.48 (s)
C-2'	75.90 (d)
C-3'	36.12 (d)
C-4'	24.34 (t)
C-5'	11.47 (q)
C-6'	15.40 (q)
HCO	161.48 (d)
OMe	53.00 (q)
C-Me	26.68 (q)
	16.55 (q)
	13.14 (q)
CH_3COO	169.71 (s)
	168.79 (s)
	167.53 (s)
CH_3COO	22.04 (q)
	21.20 (q)
	20.61 (q)

Table 3.2 ^{1}H NMR Data for nymania-1 acetate (**II**):

Proton number	
H-1	5.17 (dd)
H-9	4.27 (d, 7.7Hz)
H-11	5.27 (dd)
H-12	5.90 (d, 11.5Hz)
H-16A	2.90 (m)
H-16B	2.28
H-17	3.95 (t, 18Hz)
H-21A	7.30
H-22A	6.20
H-23	7.30
H-29A	4.08 (d, 7.7Hz)
H-29B	3.7 (dd)
2H-30	5.98 ; 5.92
CO_2Me	3.65
OAc	2.05
	1.99
	1.95
CMe	1.38
	1.20
	0.92
HCO	7.95
H-2'	4.50 (d, 4Hz)
3H-5'	0.69 (t, 9Hz)
3H-6'	0.75 (d, 7Hz)

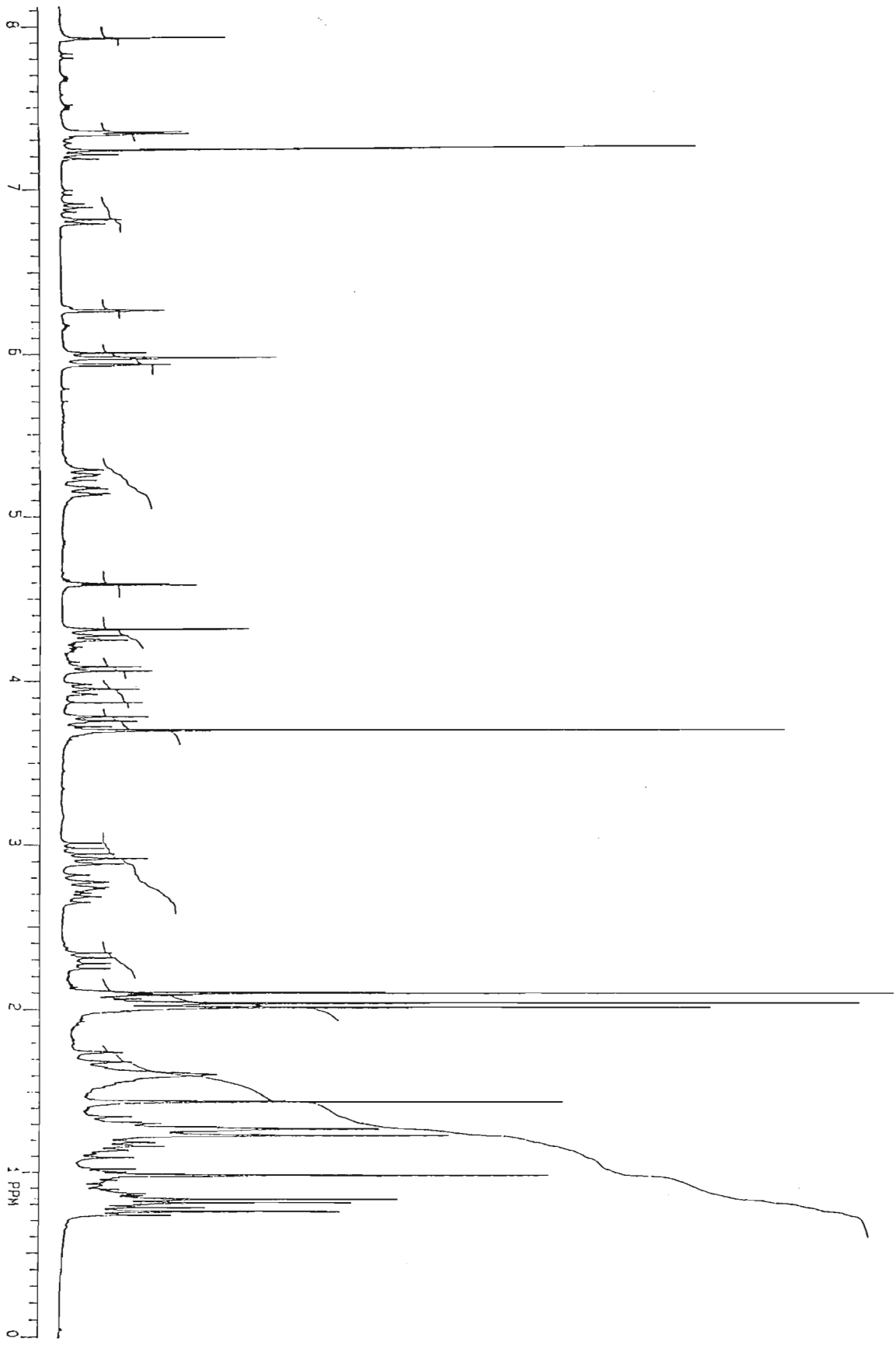


Figure 3.4 The ^{1}H NMR Spectrum of Nymania-1 acetate (**II**)

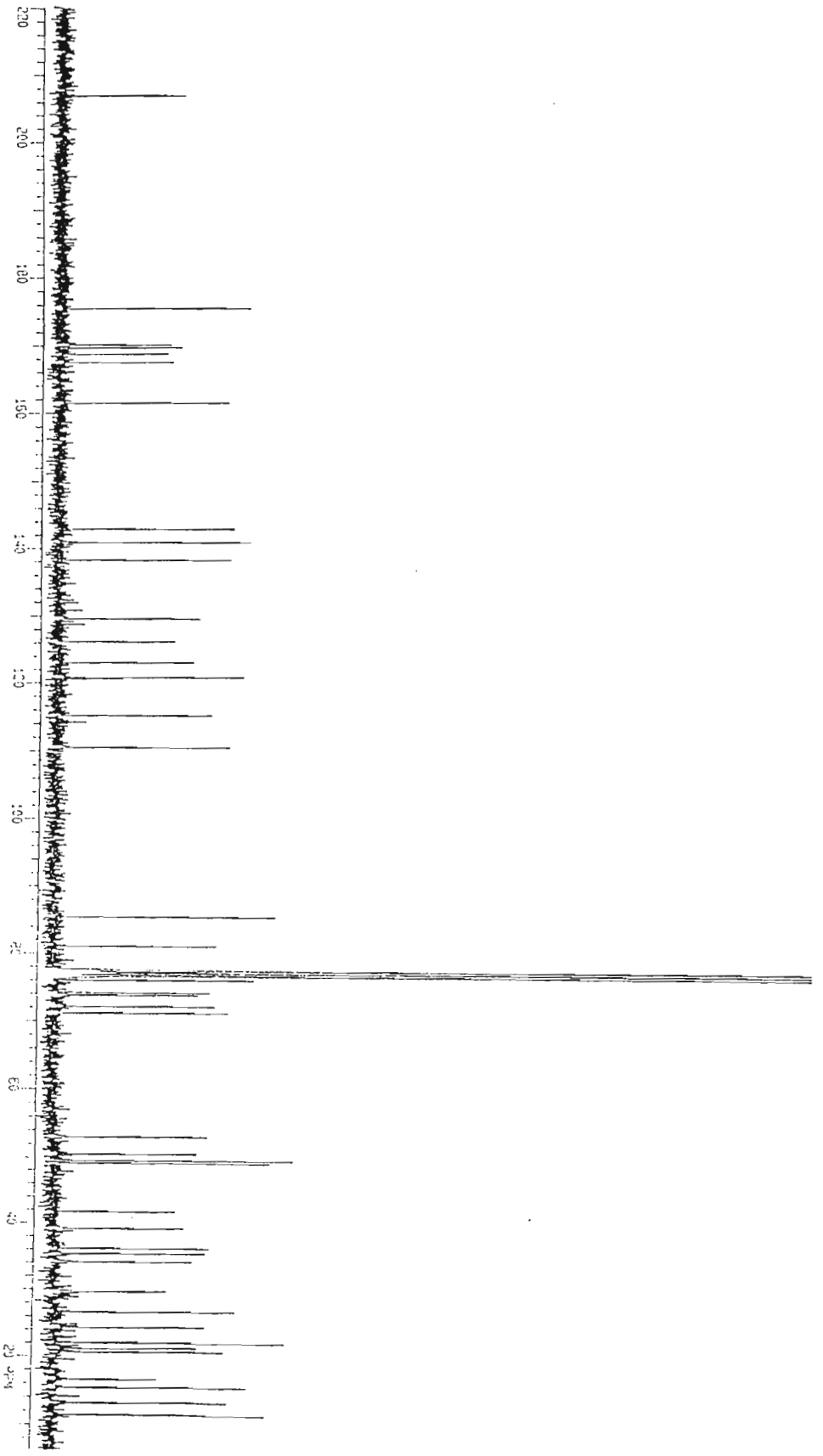


Figure 3.5 The ^{13}C NMR Spectrum of Nymania-1 acetate (II)

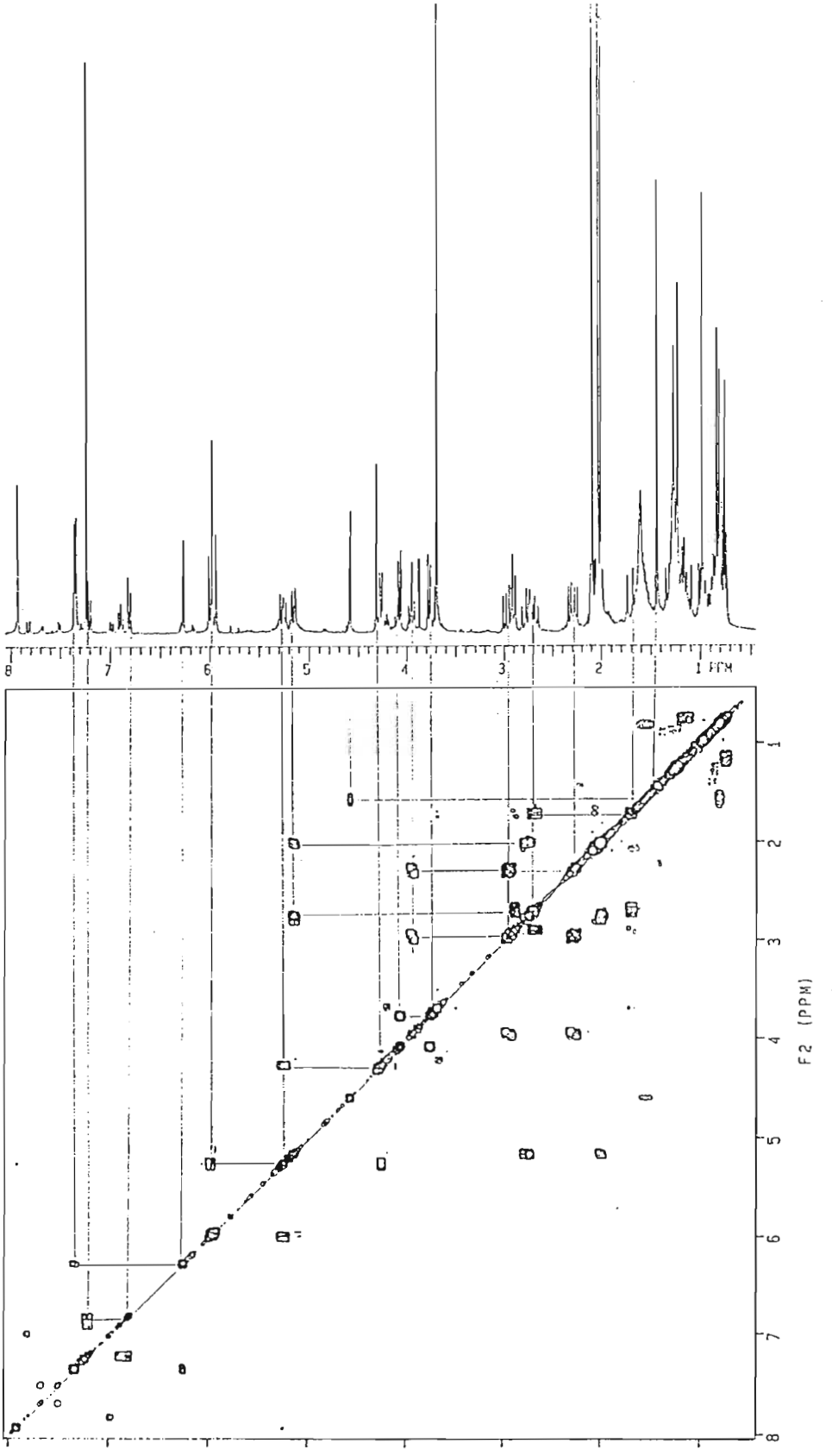


Figure 3.6 The COSY Spectrum of Nymania-1 acetate (**II**)

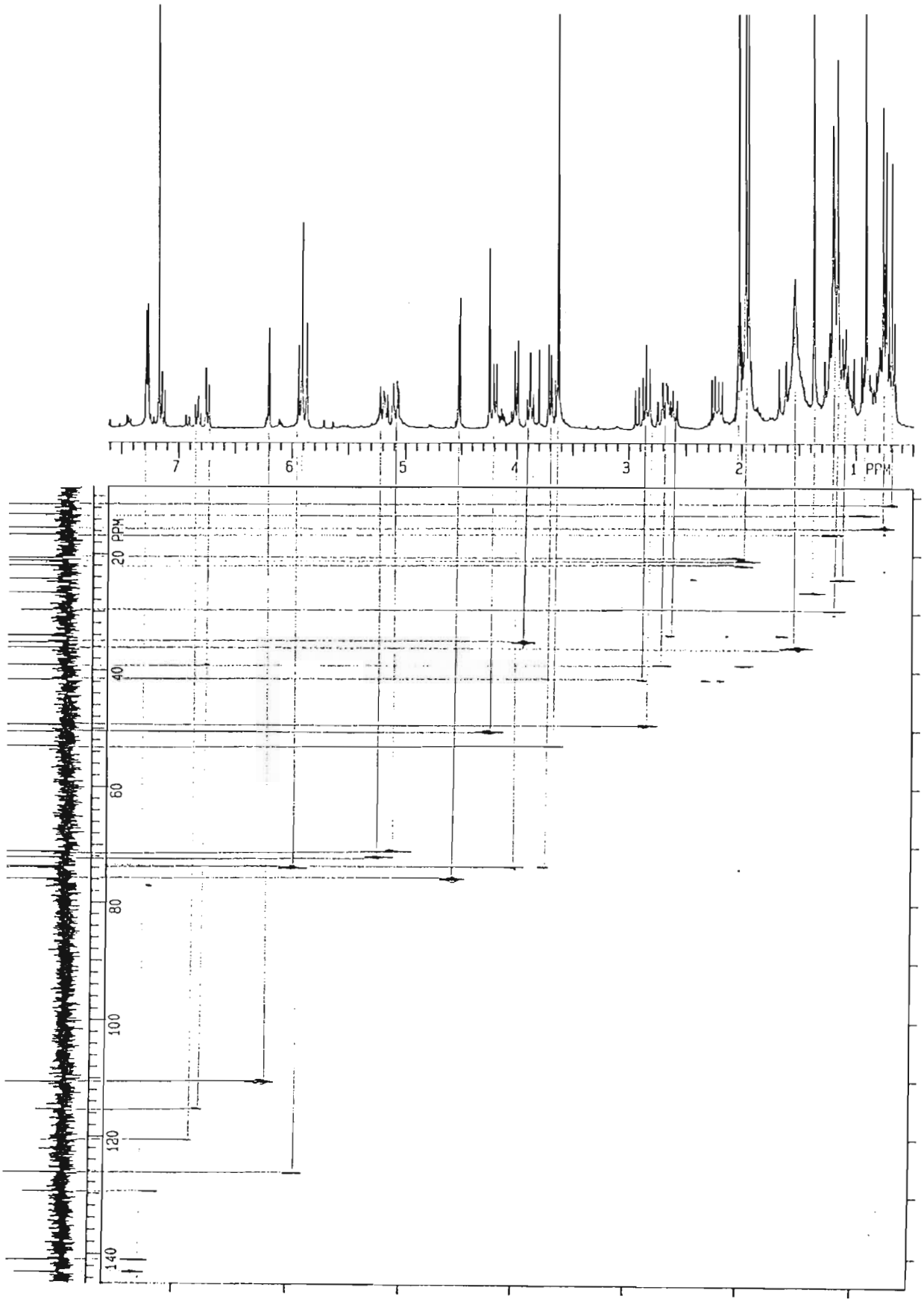


Figure 3.7 The HETCOR Spectrum of Nymania-1 acetate **(II)**

3.5.3 Rohitukin type limonoid

The ring A lactone ^{13}C singlet at δ 169.56, the formate proton resonance at δ 7.70 in the ^{1}H NMR spectrum, the triplet at C-30 (δ 124.97) indicating ring B opened and the ring D ketone ^{13}C resonance at C-15 (δ 206.47) again suggested that compound **III** was a prieurianin-type of limonoid. However, while it lacked the carbomethoxy group (methyl ester at C-7) as well as one of the acetate groups, it did have the 2-hydroxy-3-methylvalerate ester side chain at C-12. The ^{13}C NMR spectrum showed resonances which were attributable to a formate group at δ 160.37 (d, C-11), four esters or lactones at δ 174.7 (s, C-7), δ 173.05 (s, C-1'), δ 169.56 (s, C-3'), δ 169.46 (s, OAc), two ethylenic carbons at δ 124.97 (t, C-30) and δ 138.45 (s, C-8), a keto group at δ 206.47 (s, C-15), two hydroxy groups, and a furan ring. This accounted for all the oxygen atoms in the molecule. Comparison of ^{1}H and ^{13}C NMR spectra showed that the structure was almost identical to rohitukin having the 7,29 lactone, but differed in the ester attached at C-12. Rohitukin was originally isolated from *Aphanamixis polystacha* (Connolly *et al.*, 1979) and this is the first reporting of the presence of rohitukin (isolated with a different side chain) in *Turraea obtusifolia.*

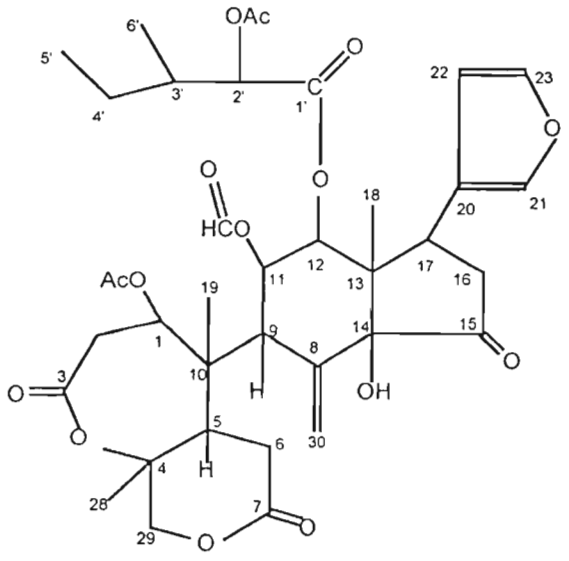


Figure 3.8 rohitukin type limonoid (**III**)

Table 3.3 ^{13}C NMR Data for rohitukin type limonoid (**III**):

Carbon atom number	
C-1	77.32 (d)
C-2	37.24 (t)
C-3	169.56 (s)
C-4	80.69 (s)
C-5	42.37 (d)
C-6	31.86(t)
C-7	174.70 (s)
C-8	138.46(s)
C-9	51.33 (s)
C-10	49.59 (s)
C-11	71.12(d)
C-12	73.09 (d)
C-13	45.82 (s)
C-14	79.42 (s)
C-15	206.47 (s)
C-16	41.67 (t)
C-17	34.97 (d)
C-20	122.68 (s)
C-21	140.60(d)
C-22	110.42 (d)
C-23	143.21 (d)
C-28	22.81 (q)
C-29	77.94 (t)
C-30	124.97 (t)
C-1'	173.05(s)
C-2'	75.58 (d)
C-3'	37.92 (d)
C-4'	22.87 (t)
C-5'	11.38 (q)
C-6'	15.27 (q)
Acetate C=O	169.46 (s)
Formate	160.37 (d)
OAc	21.26 (q)
C-Me	15.13 (q)
	12.83 (q)
	11.48 (q)

Table 3.4 ^{1}H NMR Data for rohitukin type limonoid (**III**):

Proton number	
H-1	4.92 (m)
H-2	3.24 (bs)
H-9	3.74 (d, 7Hz)
H-11	5.35 (m, 19Hz)
H-12	6.11 (d, 11Hz)
H-16	12.7 (m)
H-17	3.90 (m, 18Hz)
H-29A	4.14 (d)
H-29B	4.11 (d)
H-30A	5.83
H-30B	5.45
H-23	7.33
H-21	7.20
H-22	6.20
Ac	2.02
HCO	7.70
H-2'	3.04
H-3'	1.38
H-4'	2.75
H-5'	0.69 (t, 9Hz)
H-6'	0.77 (d, 7Hz)
CMe	1.76 (x2)
	0.90 (x3)

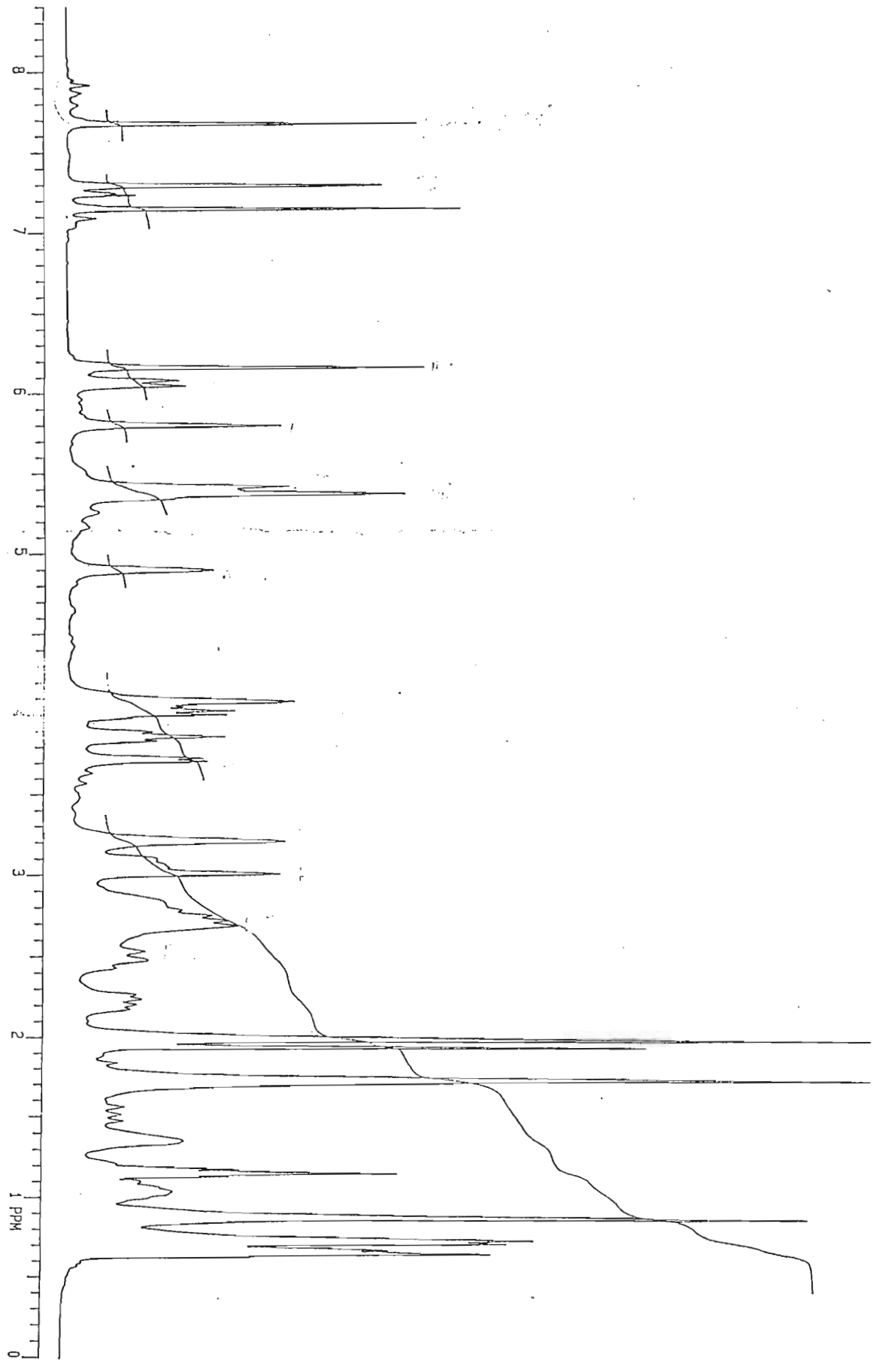


Figure 3.9 ^{1}H NMR Spectrum of rohitukin type compound

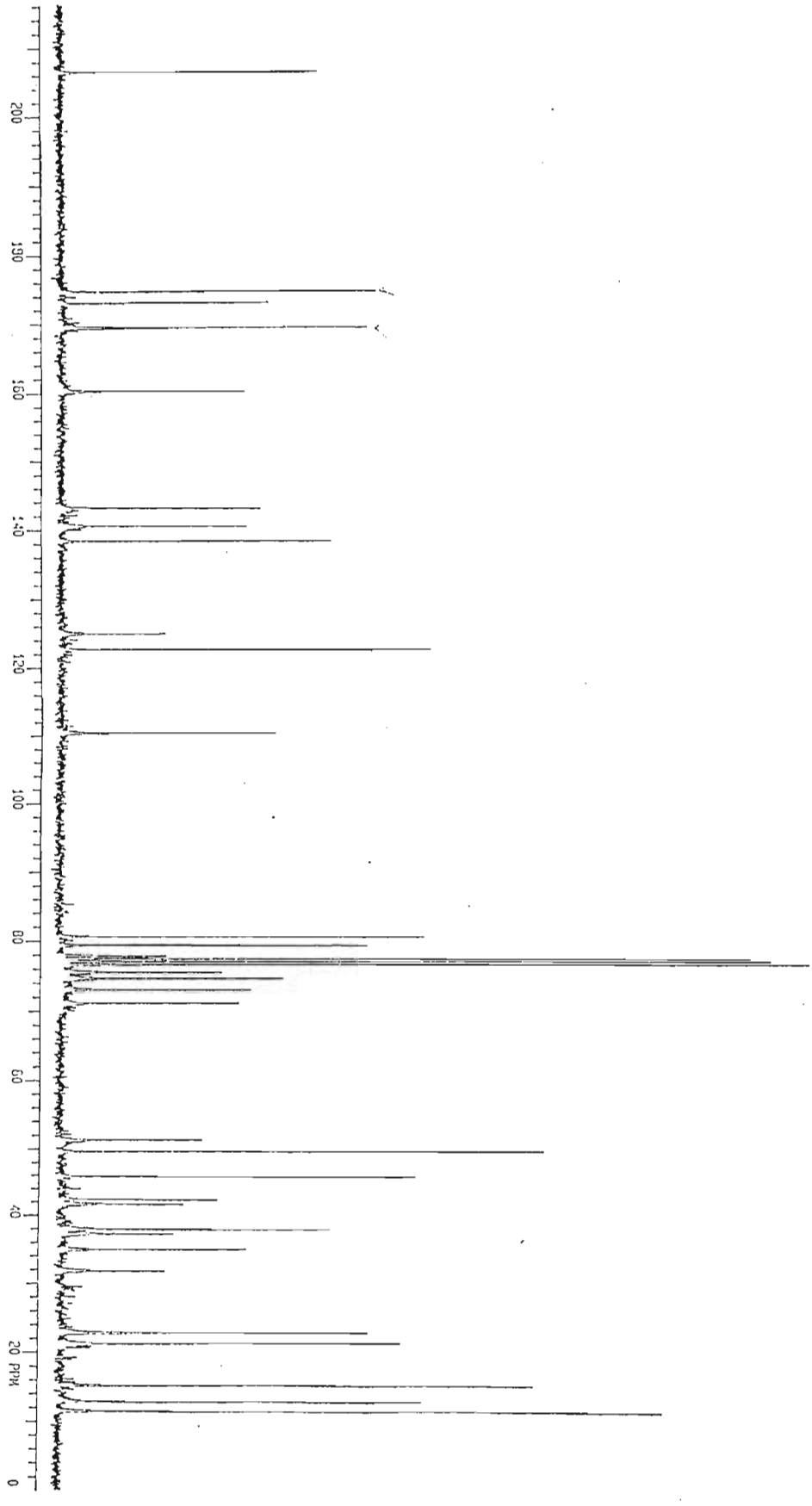


Figure 3.10 ^{13}C NMR Spectrum of rohitukin type compound

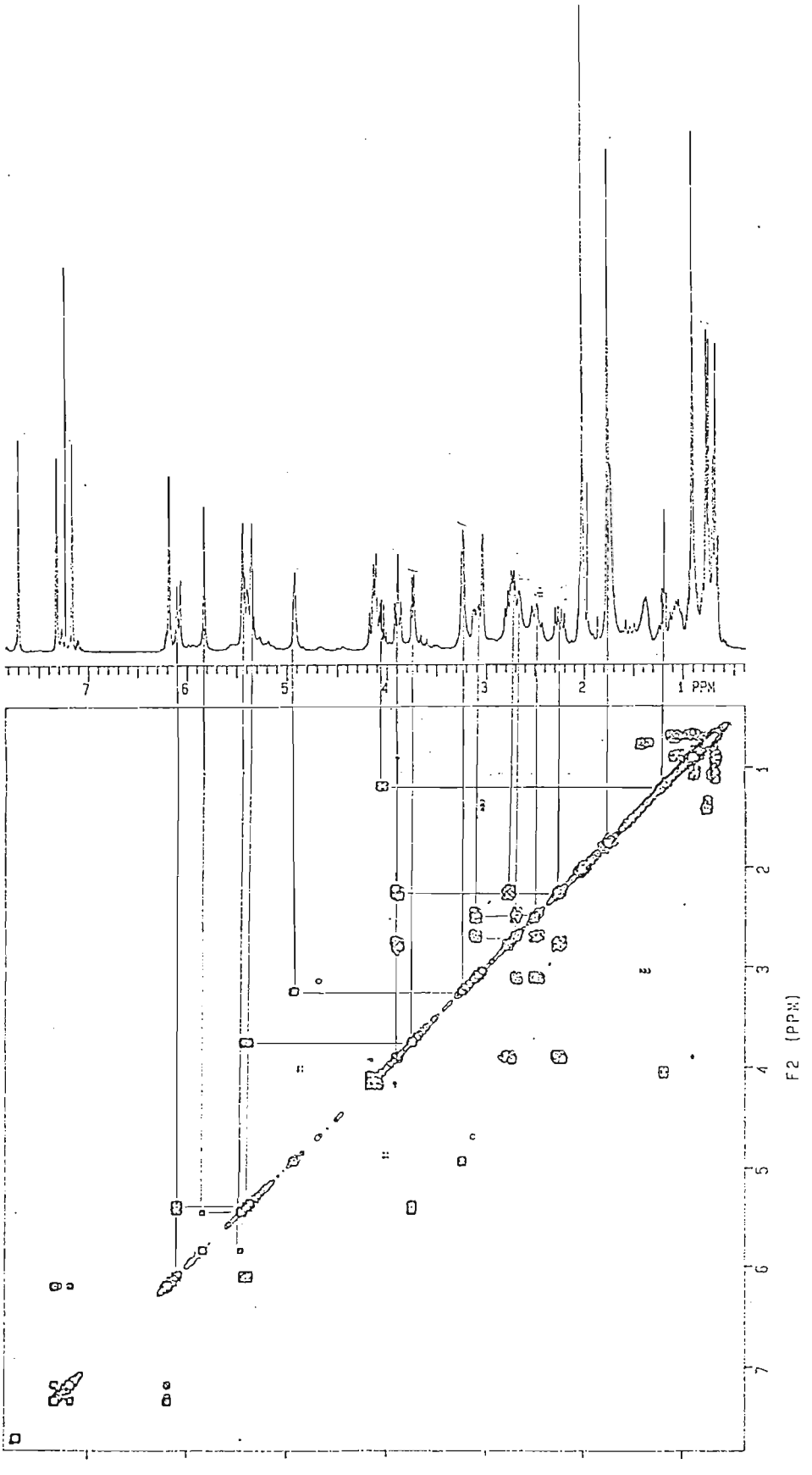


Figure 3.11 COSY Spectrum of rohitukin type compound

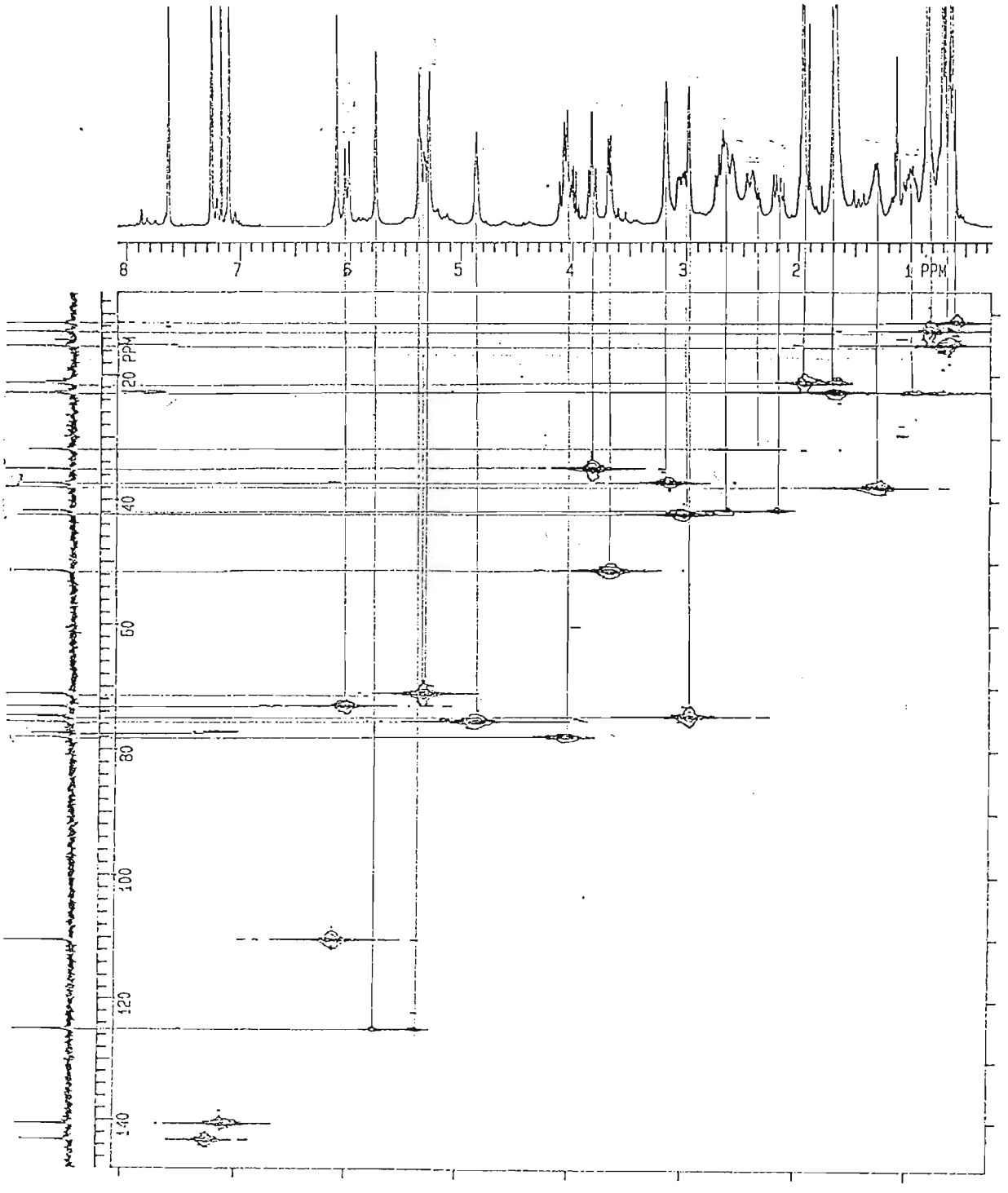


Figure 3.12 HETCOR Spectrum of rohitukin type compound

3.5.4 Chemotaxonomic Significance of nymania-1

Nymania was recognised both by De Jussieu, (1789) and Ventenat, (1799) as belonging to the Meliaceae. However, not all subsequent authors have agreed with this conclusion, and at times it has been placed in six different families. This is mainly on account of its fasciculate, sclerophyllous leaves, almost free staminal filaments and large inflated capsules which are so different in appearance from other South African members of the Meliaceae family (Sonder, 1860; Dyer, 1975). However, modern authors consider evidence of the relationship with the Meliaceae to be convincing (Pennington and Styles, 1975). Maclachlan and Taylor, (1982) showed that the bark and timber of *Nymania capensis* contained the known limonoid prieurianin (**26**), together with four other complex limonoids which had not previously been found elsewhere. These were named as nymania substances 1-4. Prieurianin (**26**) is a highly characteristic marker substance, which is widely distributed in the sub-family Meliodeae of the Meliaceae (Taylor, 1984) having been found in *Trichilia prieuriana* (Bevan *et al.*, 1965), *Guarea guidonia* (Lucacova *et al.*, 1982) and *Ekebergia pterophylla* (Mulholland, unpublished work). This provided good chemotaxonomic evidence for including *Nymania* in the sub-family Meliodeae.

However, not enough was yet known of the chemistry of the tribe Turraeeae to comment on the relationship of *Nymania* to this tribe. But since *Turraea obtusifolia* was found to contain prieurianin (**26**), a relationship between *Nymania* and Turraeeae was suggested (Akinniyi *et al.*, 1986). Pennington and Styles, (1975) placed *Nymania* in the tribe Turraeeae of the Meliodeae sub-family. The results of this research which show the presence of nymania-1 (**I**) in *Turraea obtusifolia* seeds confirms the placing of the genus *Nymania* in the tribe Turraeeae.

3.6 Summary

This chapter briefly discusses the elucidation techniques used to assign the correct chemical structure to a new compound that is essential to the establishment of reliable classification databases. The chapter also reveals the intricacies of limonoid elucidation and why information derived from ^{13}C NMR spectra was used as the basis for classification. Finally, the significance of the Meliaceae family was conveyed, and the results of recent extractive work by the author were included.

Chapter 4

Classical Classification Techniques

4.1 Introduction to Classification Modelling

Structural elucidation involves the derivation of a chemical structure from the results of various analytical techniques performed on a previously unknown chemical sample. However, classification involves the assigning of already elucidated chemical structures into categories, or "classes". This may involve determining whether or not a sample is good or bad, or predicting an unknown sample as belonging to one of several distinct groups. A classification model is used to predict a sample's class by comparing the sample to a previously analysed experience set, in which categories are already known. Classification databases can be very useful for the inexperienced researcher.

4.2 Classification Techniques

Classification by pattern recognition as in the case of ^{13}C NMR spectra may be of two types:

- unsupervised analysis
- supervised analysis

4.2.1 Unsupervised Analysis

"Although the human brain is excellent at recognising and classifying patterns and shapes, it performs less well if an object is represented by a numerical list of attributes, and much of the data is presented in that form. In the case of unsupervised pattern recognition, often referred to as cluster analysis or numerical taxonomy, no class knowledge is available and no assumptions need be made

regarding the class to which a sample may belong. Cluster analysis is a powerful investigative tool, which can aid in determining and identifying underlying patterns and structure in tables of apparently meaningless data" (Adams, 1995).

"It is implicit in producing clusters that such a group can be represented further by a typical element of the cluster. One common method of identifying a cluster's typical element is to substitute the mean values for the variates describing the objects in the cluster. The between-cluster distance can then be defined as the Euclidean distance, between these means. The nearest-neighbour distance defines the distance between two closest members from different groups while the furthest-neighbour distance is that between the most remote pair of objects in two groups" (Adams, 1995). Many numerical clustering techniques have been developed, some of which are:

4.2.1.1 Hierarchical techniques
4.2.1.2 Principal Component Analysis (PCA)
4.2.1.3 Iterative algorithms eg: K-means algorithm
4.2.1.4 Fuzzy Cluster Analysis

4.2.1.1 Hierarchical techniques

"Here the elements or objects are clustered to form new representative objects. These techniques are very popular because the application leads to the production of a dendogram which can provide a two-dimensional pictorial representation of the clustering process and the results" (Adams, 1995).

Hierarchical Cluster Analysis (HCA)

In HCA, distances between the samples (or variables) in a data set are calculated and compared. When the distances between samples are relatively small, this implies that the samples are likely similar, at least with respect to the measurements taken. Dissimilar samples will have larger relative distances. Known in biological sciences as numerical taxonomy, hierarchical cluster analysis allows the grouping of data into clusters showing similar attributes.

"The primary purpose of HCA is to present data in a manner that emphasises the natural groupings in that data set. In contrast with analytical techniques that attempt to group new samples into pre-existing categories, HCA seeks to define those categories in the first place. The presentation of HCA results in the form of a dendrogram makes it possible to visualise clustering such that relationships can be more readily seen" (Adams, 1995).

4.2.1.2 *Principal Component Analysis (PCA)*

"PCA is designed to provide one with the best possible view of the variability in a multivariate data set. This view allows one to see the natural clustering in the data, identify outliers (i.e., unusual samples) and find the reasons behind any pattern that is observed. In addition, the intrinsic dimensionality of the data can be determined and, with variance retained in each factor and the contribution of the original measured variables to each, this information can be used to assign chemical meaning (or biological meaning or physical meaning) to the data patterns that emerge and to estimate what portion of the measurement space is noise.

"PCA is fundamentally similar to factor analysis or eigenvector analysis (mathematics) or even perceptual mapping (marketing). It is a method of transforming complex data into a new perspective in which, hopefully, the most important or relevant information is made more obvious. This is accomplished by constructing a new set of variables that are linear combinations of the original variables in the data set. These new variables, often called eigenvectors or factors, can be thought of as a new set of plotting axes which have the property of being orthogonal (i.e., completely uncorrelated) to one another" (Adams, 1995).

4.2.1.3 *Iterative algorithms eg: K-means Algorithm*

"Here methods are employed which optimise the partitioning between clusters using some type of iterative algorithm, until some predefined minimum change in the groups is produced. The objective of the method is to partition the m objects, characterised by n variables, into K-clusters so that the square of the within-cluster sum of distances is minimised. Being an optimisation-based technique, the number of solutions cannot be predicted and the best possible partitioning of the objects may not be achieved" (Adams, 1995).

4.2.1.4 Fuzzy Cluster Analysis

"Here objects are assigned a membership function indicating their degree of belonging to a particular group or set. Unlike the previously mentioned techniques which reveal nothing about how well any specific object fits into its chosen cluster, or how close it may be to other clusters, this method not only highlights similar objects but also provides information regarding the relationship of each object to each cluster" (Adams, 1995).

These four techniques described above would be of little value in our research as ours was essentially a classification problem in which the target output goals were predefined. Thus it would be pointless to employ any of these techniques, as, in so doing, useful information would be discarded.

4.2.2 Supervised Analysis

"Supervised pattern recognition is often referred to as classification or discriminant analysis. The number of parent groups is known in advance and representative samples of each group are available. The means of deriving the classification rules from previously classified samples is referred to as discrimination. Thus, a suitable collection of pre-assigned objects, the training set, must be available to determine the discriminating rule. If the parent population distribution follows the normal curve, then parametric methods such as statistical discriminant analysis can be used. If the distribution of the data is known not be normal, then non-parametric methods are used. One of the most widely used non-parametric algorithms is that of K-nearest neighbour" (Adams, 1995).

4.2.2.1 Bayes' Theorem

"The Bayes' rule states that "a sample or object should be assigned to that group having the highest conditional probability (i.e.: the probability of an event is modified, for better or worse, by prior knowledge)" and application of this rule to parametric classification schemes provides optimum discriminating capability" (Adams, 1995).

4.2.2.2 Nearest Neighbours

"The discriminant analysis techniques rely for their effective use on an *a priori* knowledge of the underlying parent distribution function of the variates. However, this is often not the case and one of the most common non-parametric techniques is known as K-nearest neighbours.

"The general method is based on applying the K-nearest neighbour classification rule, referred to as K-NN. The distance between the pattern vector of the unclassified sample and every classified sample from the training set is calculated, and the majority of smallest distances, ie: the nearest neighbours, determines to which group the unknown is to be assigned. The most common distance metric used is the Euclidean distance between two pattern vectors" (Adams, 1995). In other words, KNN is based on a distance comparison among samples: an N-dimensional distance between all samples in the data set is calculated, where N is the number of variables in the measured data. The predicted class of a test sample is then determined based on the identity of those samples closest to the unknown sample. This is accomplished in a fashion analogous to voting: each of the K nearest samples votes once for its class; the class receiving the most votes is assigned to the test sample. "K-NN is a multi-category method that does not require repeated application to assign some unknown sample to a class as is often the case with binary classifiers" (Adams, 1995). It is also tolerant of sample-poor situations and is the only method that works well even when categories are strongly sub-grouped.

"Its major disadvantage is that it is computationally demanding. For each classification decision, the distance between the sample pattern vector and every object in the training set for all groups must be calculated and compared. Thus, when very large data sets are used, each distinct class or group can be represented by a few representative patterns to provide an initial first-guess classification before every object in the best classes is examined" (Adams, 1995).

4.2.3 Advantages and Disadvantages of Analysis

Advantages:

- rigorous
- deterministic/explicit
- statistically significant results

Disadvantages:

- rely on probabilistic methods
- cannot learn from real data
- produce hard-decision outputs
- do not accommodate noisy data well (Haykin, 1994)
- some require *a priori* knowledge of the classification groups
- are not well suited to non-linear problems

4.3 Human classification

Humans are naturally skilled at pattern recognition and can learn relevant characteristics of phenomena. However, human accuracy is sometimes questionable and biased by experience. Humans also find detection of multiple high dimensional relationships difficult to analyse and can be distracted by noise. Similarly, experientially gained knowledge is difficult to quantify or pass on to others. In structure elucidation from NMR spectra, people require many years to reach competence.

4.3.1 Advantages and Disadvantages

Advantages:

- good at pattern recognition
- can learn relevant characteristics
- knowledge of biochemical pathways can positively influence the decision

Disadvantages:

- accuracy is questionable
- human knowledge is biased by experience etc.
- poor repeatability
- difficult to detect hidden relationships
- distracted by noise
- cannot easily pass on information
- requires many years of exposure to correctly classify

4.4 Expert systems

Expert systems are a technique of extracting expert human knowledge and writing it as a series of sequential rules. Based on these rules novel data is filtered through a decision tree structure. The knowledge resides in a knowledge base which is a repository of information relating to the problem. The advantage of expert systems is that human expert knowledge is easily captured and that the decision process is explicit at every decision step (or branch of a decision tree). They are limited, however, by incomplete or biased knowledge and those exceptions to each decision rule need to be written which are not always intuitive. Furthermore the decision space covered by an expert system is frequently either too specific or if general enough, too coarse to recognise fine detail. Expert systems store knowledge in syntactical format that makes it nearly impossible to learn from real data.

4.4.1 Advantages and Disadvantages

Advantages

- human knowledge is easily captured
- the decision process is explicit at every decision step (or branch of the tree)

Disadvantages

- frequently the knowledge used to build the knowledge-base is biased or incomplete
- exceptions to each decision rule need to be written (not always intuitive)

- the decision space covered by an expert system is frequently either too specific or if general enough, too coarse to recognize fine detail
- the system cannot learn from real data/experimental data
- rules are performed sequentially, whereas most natural decisions are highly parallel in nature

4.5 Summary and Conclusion

In this chapter, a brief overview of automated classification techniques as well as human classification is discussed. The extractives from the Meliaceae family have very complex ^{13}C NMR spectra. In addition to their large molecular size (30 to 40 carbon atoms), they are skeletally very intricate and structurally similar. Some structures differ only by one functional group eg: those compounds that have a hydroxyl group replaced by an acetate group. Thus the spectra of the two compounds differ only by three carbon resonances out of approximately 35. This difference, compounded with extraneous peaks associated with natural product isolation, becomes even more insignificant. Other inherent problems associated with natural product isolation include:

- very small amounts of sample which produce very weak ^{13}C NMR signals
- a resultant low signal-to-noise ratio and superposition of peaks
- samples which are never absolutely pure and which are also often isolated as inseparable stereoisomeric mixtures or mixtures of esters
- measurement of line intensities is non-quantitative and non-reproducible due to the Nuclear Overhauser effect and Fourier Transform pulse intervals

These characteristics make compound classification using the classical computational techniques in this discussion very difficult. The advantage of human generalisation as well as computational speed and accuracy are required, and thus another approach was sought.

Chapter 5

Introduction to Neural Networks

The three characteristics that describe a neural network are structure, dynamics (falls into meso-structure) and learning.

A key strength of neural networks is the ability to adaptively learn. Unlike traditional expert systems that could only identify an exact manifestation of a pattern in its database, neural networks discriminate between patterns based on examples and training. An example of this is the well-known experiment in which neural networks learned the difference between convex, concave and straight lines, regardless of whether the lines were shifted, the curve inflection changed or noise added. Remarkably the neural networks could correctly determine the line type when only a portion of the curve was presented for analysis.

In constructing a neural network, the principal focus is the design of the architecture and not how the network will learn to discriminate among the possible choices. The designer must develop a good learning algorithm that will let the network learn to discriminate. Some networks continue to learn after their initial training period is completed and hence they can modify their categories if the types of examples change.

Neural networks use their adaptive learning capabilities to self-organise the information they receive during learning. When the network self-organises, it creates representations of distinct features in the presented data. Even when networks are taught to recognise certain classes of patterns, they self organise the information used for pattern recognition e.g.: the back-propagation network will create its own feature representation by which it recognises certain patterns. This leads to generalisation that allows the network to respond appropriately when confronted with non-uniform data.

Today is an exciting period of transition for neural networks. The primary applications of back propagation networks are pattern recognition, signal processing and classification.

5.1 Introduction

"Artificial neural networks have evolved from mainstream artificial intelligence as an attempt at modelling learning in biological neural systems. Whereas rule-based expert systems, fuzzy logic and case-based reasoning rely on logical syntax for decision-making (syntactical artificial intelligence), neural networks depend on a more abstract symbolic representation of data. Whereas traditional artificial intelligence rests heavily on a rigorous set of IF-THEN rules, historical precedents or partial set membership, neural networks attempt to learn the underlying relationships present in data without such formalism. They rely on the recognition of patterns latent in the data without being explicitly shown which patterns or which inputs to use in the association.

"There are several distinguishing features of neural computing which differs from other artificial intelligence methods and more traditional techniques:

Learning by example

"Neural networks generate their own rules by learning from examples that they have been shown.

Distributed associative memory

"Neural networks store their information in a distributed way. This leads to an important capability of neural networks namely generalisation of input data. Generalisation allows "intelligent" responses to novel stimuli.

Fault tolerance

"Unlike traditional systems that are rendered useless by even a small amount of damage to memory, neural networks are much more robust and fault tolerant. Fault tolerance extends to graceful degradation with increased neural network corruption.

Relative noise immunity

"The generalising ability of neural networks allows a certain amount of tolerance to noisy, missing or incomplete data.

Pattern recognition

"Neural networks are much more adept at pattern recognition tasks than traditional techniques. Their ability to deduce complex relationships between data gives them an edge over expert systems.

Functional synthesis

"Certain neural networks are able to learn complex continuous non-linear mappings from one or more inputs to one or more outputs.

Other advantages

"Certain neural networks are capable of adaptive learning that is useful in dynamically varying environments. Neural networks are often cheaper and quicker to develop than other artificial intelligence techniques" (Maren *et al.*, 1990). A typical development cycle for neural networks is shown in Figure 5.1.

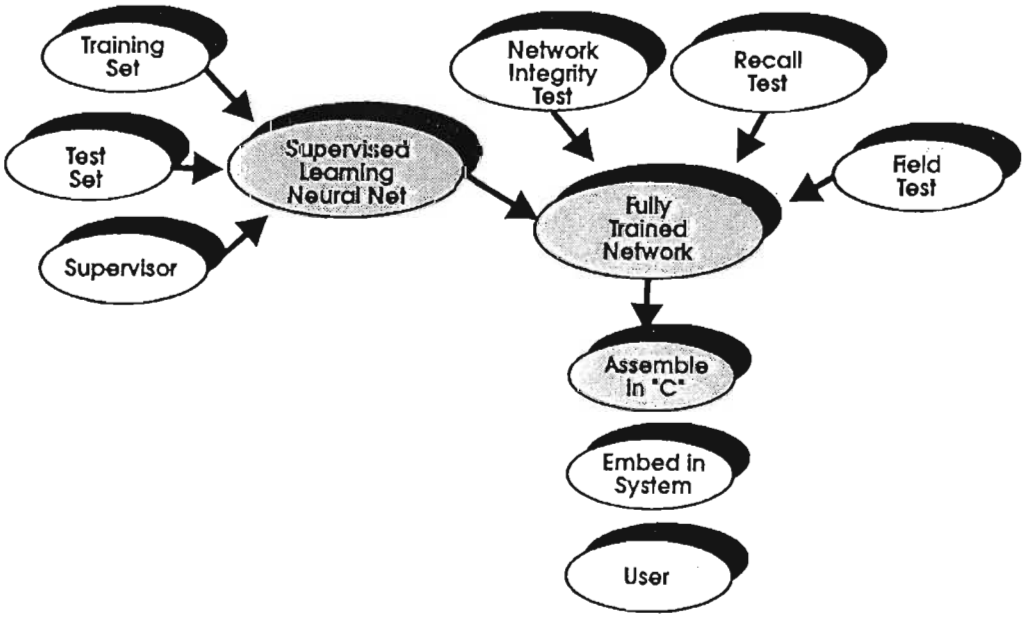


Figure 5.1 Neural Network development cycle

5.1.1 Biological neurological basis for neural networks

"Animal brains consist of billions of neurons connected in a massively parallel manner. Each neuron, for example, typically has 10 thousand connections to its neighbours. The speed of signal propagation in the connections is around 1-100 m/s, not fast by digital computing standards. Similarly, the speed of each neuron in processing incoming data is relatively low. It can calculate the weighted sum of the signal strengths on these inputs in around 5 ms. The performance of a biological brain is therefore dependent more on its parallel architecture than its elemental processing speed.

"An example of a simplified biological neuron is shown in Figure 5.2. The connections from thousands of neighbouring neurons enter through the axons that connect to the cell body. The output of the neuron is via the dendrite to its other neural neighbours. Neurons trigger or "fire" in response to the relative strength of the excitatory (+ve) or inhibitory (-ve) signals on its input. The decision whether to trigger or not is not based on a single event or a occurrence, but on the cumulative response of a network of neural elements to a stimulus. It appears that learning in biological systems takes place in the adjustment of the relative strengths of the incoming connections to each neuron; the greater the connection strength, the more important or influential that particular input is. Learning in which similar patterns strengthen connections is termed Hebbian learning after its discoverer Donald Hebb.

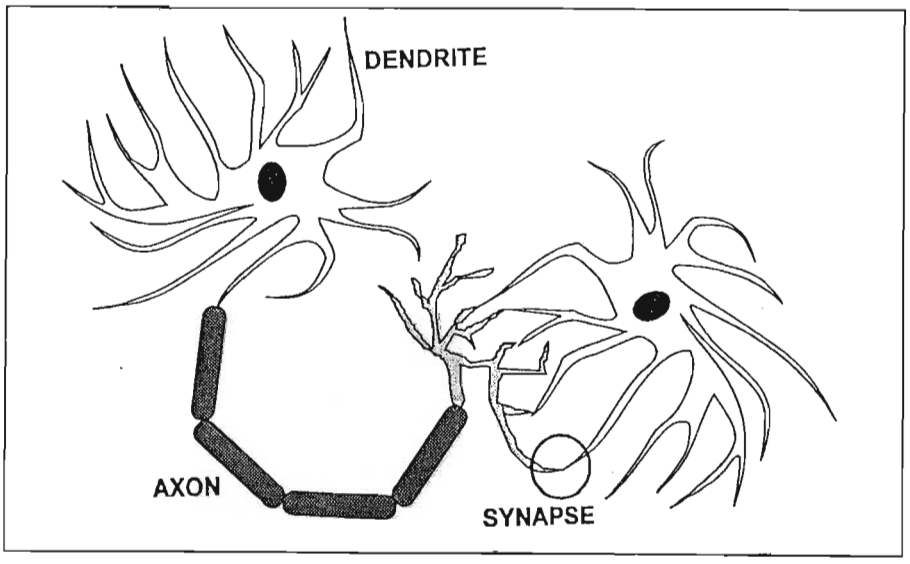


Figure 5.2 Simplified biological neuron (Maren *et al.*, 1990)

"Thus information in neural networks is stored in the connection strengths between the neurons and not in discrete locations. This distributed representation of the information learned represents a departure from localised knowledge and a greater level of knowledge abstraction than is found in other artificial intelligence methods.

"Artificial neural networks attempt to model biological neural systems in a very loose way. The biological metaphor is carried through into a multiplicity of highly interconnected artificial neurons. Each neuron provides the simple function of summing the input signals in proportion to their connection strength and passing this weighted sum through a non-linear transfer function. The processing power of artificial neural networks, like their biological prototypes, is found in their highly connected, massively parallel network structure. Artificial neural networks, however, differ considerably from biological neural networks in practical implementation" (Maren *et al.*, 1990).

5.2 Neural network structures

5.2.1 Microstructure

"The microstructural basis of neural networks is the neuron, also known as the processing element. The neuron structure used in most synthetic neural networks is given in Figure 5.3. The incoming signals, called activations are either from preceding neurons or from the data source directly. Each activation is multiplied by its connection strength, or weight, and then summed. This weighted sum is then passed through a transfer function, the output of which connects to subsequent neurons.

"TRANSFER FUNCTION - specifies how the neuron will scale its response to incoming signals, and produces the neuron's activation. If the activation passes some threshold criteria, then the artificial neuron will output a signal to the neurons to which it is connected." The LOGISTIC SIGMOID (CONTINUOUS-FUNCTION) node shall be implemented in this thesis.

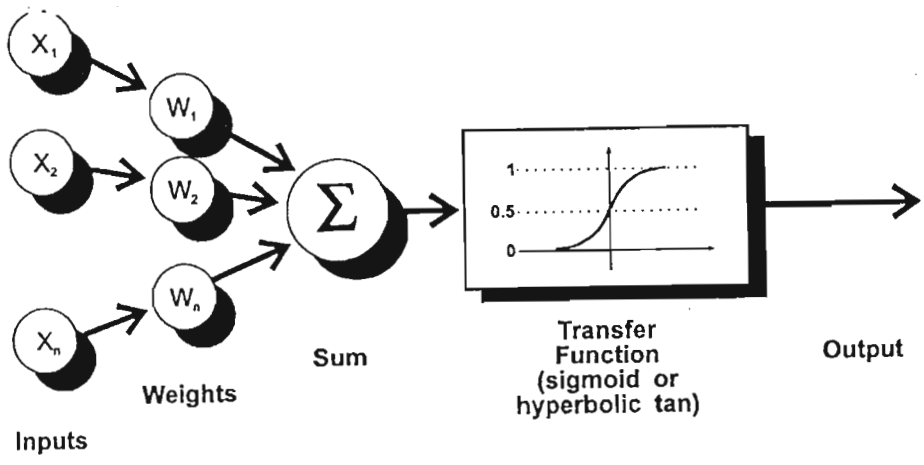


Figure 5.3 Artificial neuron microstructure (Maren *et al.*, 1990)

"A bias signal is usually added to the weighted sum to provide a fixed-level signal useful for shifting the operating point on the transfer function on all neurons equally (Maren *et al.*, 1990).

$$A_j = f\left[\sum_{i=1}^{n} w_{ij} A_i + \theta_j\right] \tag{5-1}$$

where

A_j is the output activation of neuron j

w_{ij} is the weight between neuron i and neuron j

A_i is the incoming activation of neuron i

θ_j is the bias signal for neuron j

"Several functions are useful as the neuron transfer function. Non-continuous functions are shown in Figure 5.4 (threshold logic) and Figure 5.5 (hard limit). The threshold logic is of limited use in neural networks, but the linear function with hard limits is useful in neural networks where a continuous-valued output is required, such as general regression analysis:

$$y = 0(x < -\theta) \tag{5-2a}$$

$$y = \frac{x}{2\theta} + \frac{1}{2}(-\theta \le x \le \theta) \tag{5-2b}$$

$$y = 1(x > \theta) \tag{5-2c}$$

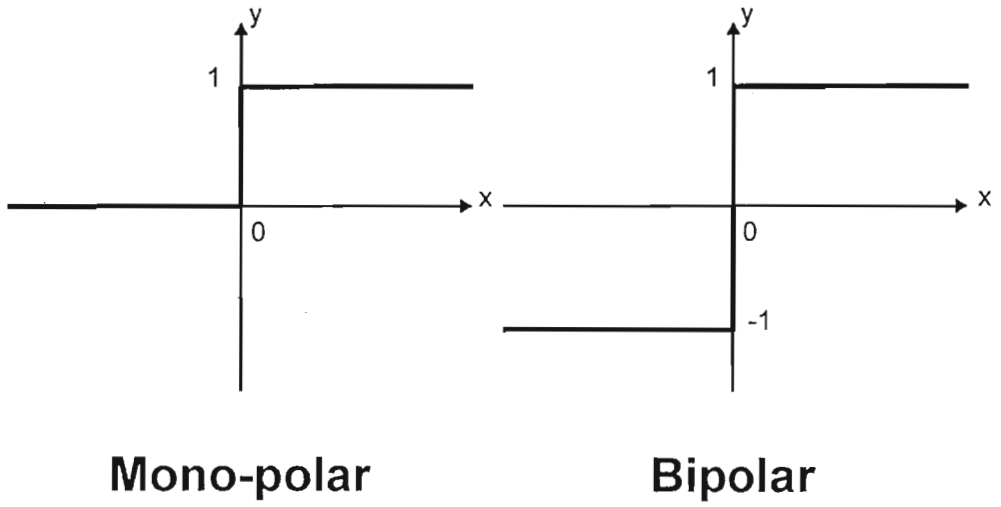


Figure 5.4 Threshold logic neuron transfer function (Maren *et al.*, 1990)

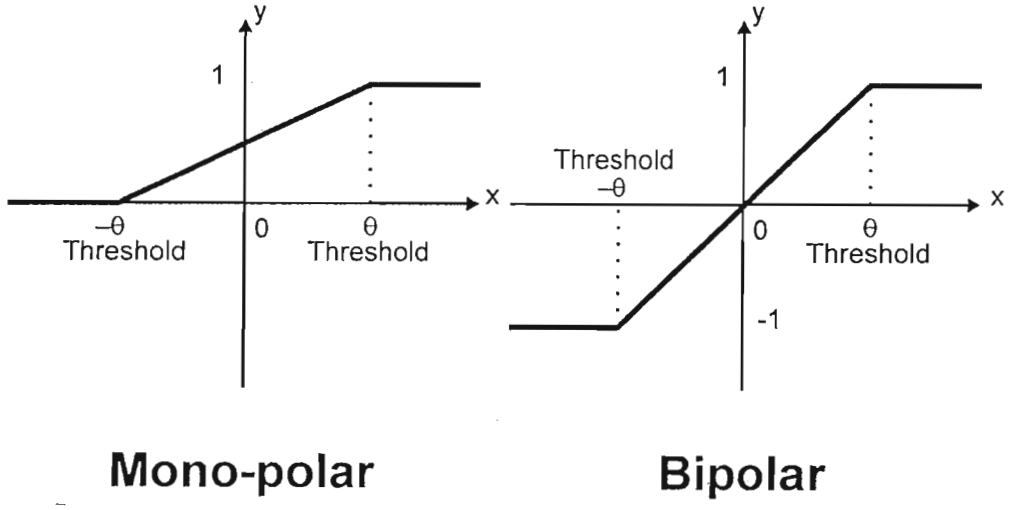


Figure 5.5 Hard-limit neuron transfer function (Maren *et al.*, 1990)

"Continuous functions are much more useful and provide non-linearity to the neurons. They are smooth over -∞ to +∞ and are in general monotonically increasing. In addition, they must be asymptotic to limits in the extremes. The two most common functions are the logistic sigmoid and the hyperbolic tangent shown in Figure 5.6 and 5.7. The logistic sigmoid is used for monopolar inputs scaled to be between 0 and 1, and the hyperbolic tangent for bipolar inputs scaled between -1 and +1:

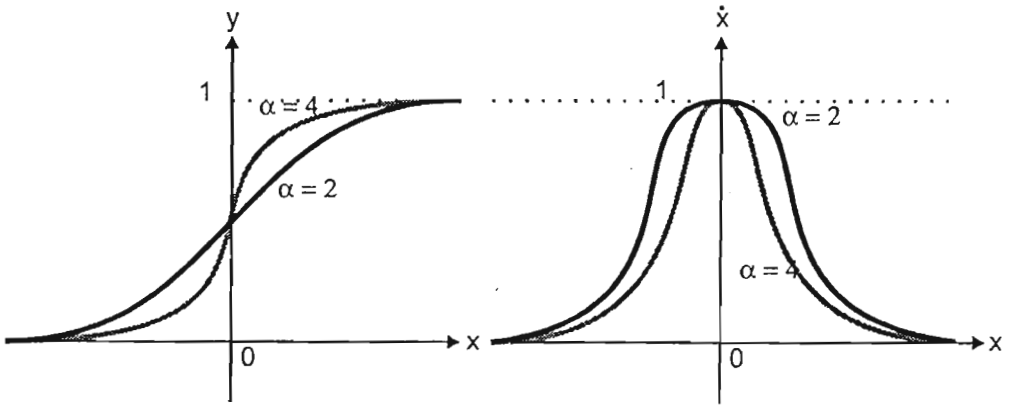
Logistic Sigmoid:

$$f(x,\alpha) = \frac{1}{1+e^{-\alpha x}} \qquad (5\text{-}3)$$

Hyperbolic Tangent:

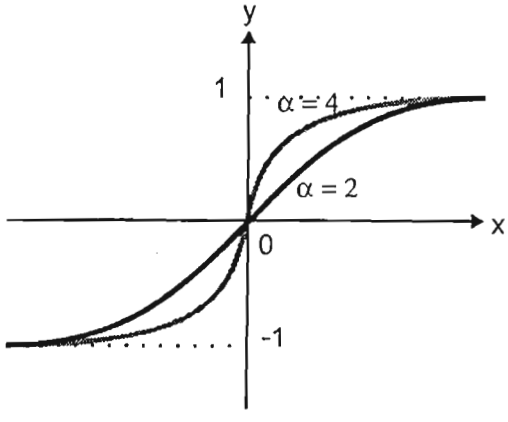
$$f(x,\alpha) = \frac{e^{\alpha x} - e^{-\alpha x}}{e^{\alpha x} + e^{-\alpha x}} \qquad (5\text{-}4)$$

where α is the slope variable



Sigmoid **Derivative**

Figure 5.6 Logistic Sigmoidal neuron transfer function and its derivative Maren *et al.*, 1990)



Hyperbolic Tangent

Figure 5.7 Hyperbolic tangent neuron transfer function (Maren *et al.*, 1990)

"For the majority of input values, the functional output is near the asymptotes, i.e. either inhibitory or excitatory. The greatest effect of the functional non-linearity is on the small range of values clustered about zero.

"The derivative of these functions yields a maximum at zero and a minimum at the extremes" (Maren *et al.*, 1990). This characteristic allows training of neurons as shown in section 5.3.

5.2.2 Mesostructure

"The meso-structure of neural networks relates to their physical organisation and the arrangement and connection of neurons. The meso-structure is the primary distinguishing feature of classes of neural network. Neural networks are classified by the following characteristics:

- number of layers (also called slabs)
- number of neurons (also called processing elements) per slab
- the type of connections (forward, backward, lateral, implicit)
- the degree of inter-neuron connectivity

"Based on these parameters, there are five major classes of structurally related neural networks.

5.2.2.1 Multilayer feedforward neural networks

"In multi-layer networks the inputs to the neural network are passed through a layer known as the input layer which merely serves to connect each input to every neuron in the subsequent layer. The input layer performs no operation on the data. The output layer is a fully functional layer of neurons whose outputs indicate the output vector in response to an input vector. Between the input and output layers there are

several layers of neurons called hidden layers. There are usually one and very rarely more than three hidden layers.

"As indicated in Figure 5.8, signals propagate in the forward direction only. There are no feedback connections for each neuron, nor lateral or backward connections. The flow of information for decision making is in the forward direction only. The most famous and widely used neural network is the back-propagation neural network. Back-propagation refers to the method of training the neural network and not the direction of information flow. It is still a feedforward neural network. General applications of feedforward neural networks include system modelling, prediction through non-linear general regression analysis, classification by pattern recognition and filtering." (Maren *et al.*, 1990) Details of the back-propagation neural network are given in section 5.3.

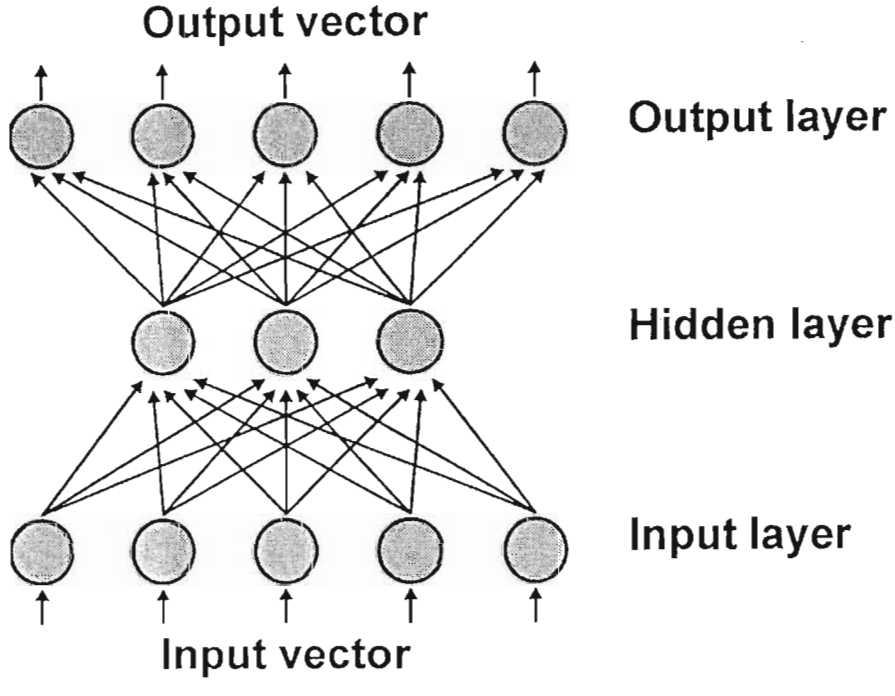


Figure 5.8 Feed-forward neural network structure (Maren *et al.*, 1990)

5.2.2.2 Single layer laterally connected neural network

"These neural networks consist of a single layer of neurons with bi-directional lateral connections (Figure 5.9). They are mainly used in auto-associative applications where the neural network outputs a stored pattern in response to the same pattern appearing at its inputs most often partially complete or noisy. If the

desired neural network output is the same as the input for all the input patterns it is termed auto-associative. If, however, the desired output is different to the input patterns it is termed hetero-associative. Single layer neural networks are sometimes provided with feedback on themselves via closed loops. This gives rise to a sub-class named recurrent neural networks. Examples of this class include the Hopfield neural network and the Brain-State-in-a-Box neural network" (Maren *et al.*, 1990).

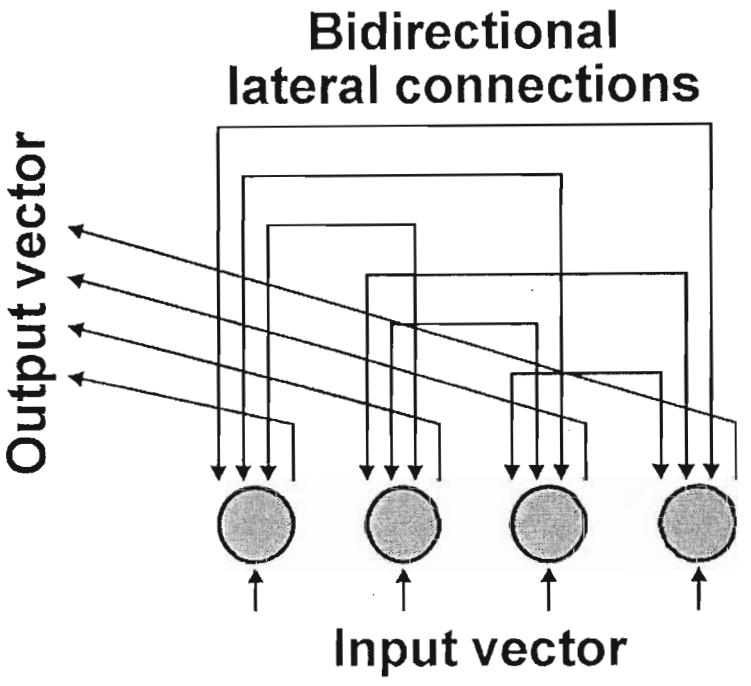


Figure 5.9 Single-layer neural network with explicit connections (Maren *et al.*, 1990)

5.2.2.3 Single layer topographically ordered neural network

"This neural network has no explicit connections as in the previous neural network and is used in cases where distinct categories of classification are already known. Inputs to the neural network are grouped as vectors called neurodes. Each neurode represents an individual from a representative selection of training vectors. They are topographically ordered with respect to one another by comparing their relative Euclidean vector distances in vector space. The shape of the neighbourhood of comparison and measures of distance are two prime parameters in designing these neural networks (Figure 5.10).

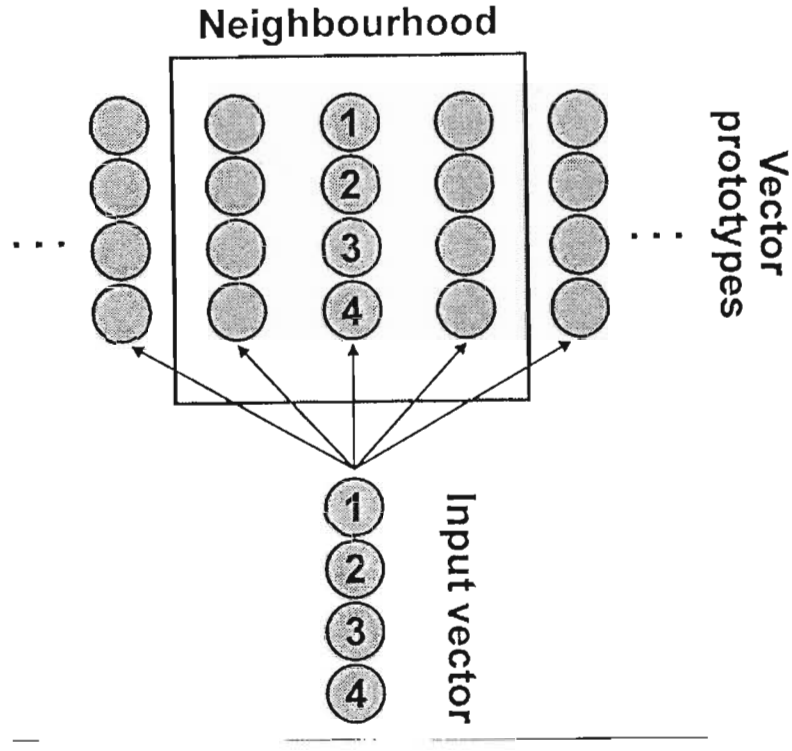


Figure 5.10 Single-layer neural network structure with implicit connections (Maren *et al.*, 1990)

"Once the neurodes are organised according to their categorical similarity or dissimilarity, a new vector, when presented to the neural network, will result in a projection into a region of vector space (category) with which it is most similar. This allows robust classification of data. Specific details of this method are contained in the section on Learning Vector Quantisation (section 5.4)" (Maren *et al.*, 1990).

5.2.2.4 *Bilayer feedforward/feedback networks*

"These neural networks contain both feedforward and feedback connections in a twin-layered structure. A simplified generic structure of these neural networks is shown below (Figure 5.11). In these neural networks, the forward and reverse connections are not just the transpose of one another but usually have different connection strengths (weights) in each direction. Because of the bi-directional flow of information in these neural networks, they are also known as resonating neural

networks. Patterns in each layer stimulate patterns in one another in a state of resonance until a stable state exists in each layer" (Maren *et al.*, 1990).

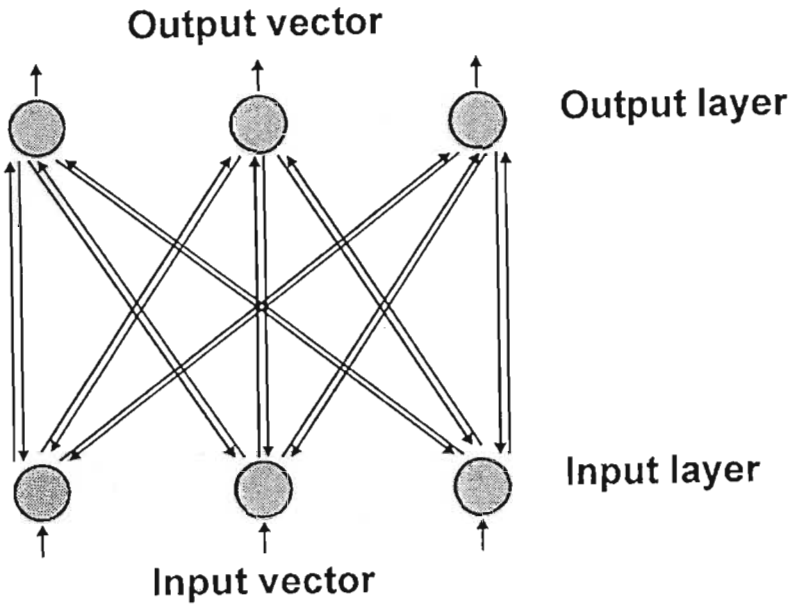


Figure 5.11 Bi-layer feedforward/feedback neural network structure (resonating) (Maren *et al.*, 1990)

The prime application of this class of neural network is in pattern recognition. Their advantage over other architectures used in pattern recognition is the ability to learn new patterns without losing memory of or degrading the recall performance of previously stored patterns. Examples of these neural networks are the Adaptive Resonance Theory (ART) networks of Carpenter and Grossberg (1987a, 1987b, 1988), and the Bi-directional Associative Memory (BAM) of Kosko (1987).

5.2.2.5 *Multilayer co-operative/competitive neural networks*

"Figure 5.12 shows schematically that these networks have lateral connections in addition to feedforward and feedback connections. The lateral connections allow individual neurons in the neural network to compete/co-operate for the opportunity of passing information. The competitive (inhibitory) and co-operative (excitatory) nature of these neural networks improves classification ability where a winner-

takes-all strategy suppresses spurious classifications and enhances most-likely classifications" (Maren *et al.*, 1990). Examples of these networks include the Probabilistic Neural Network of Specht (1988, 1990).

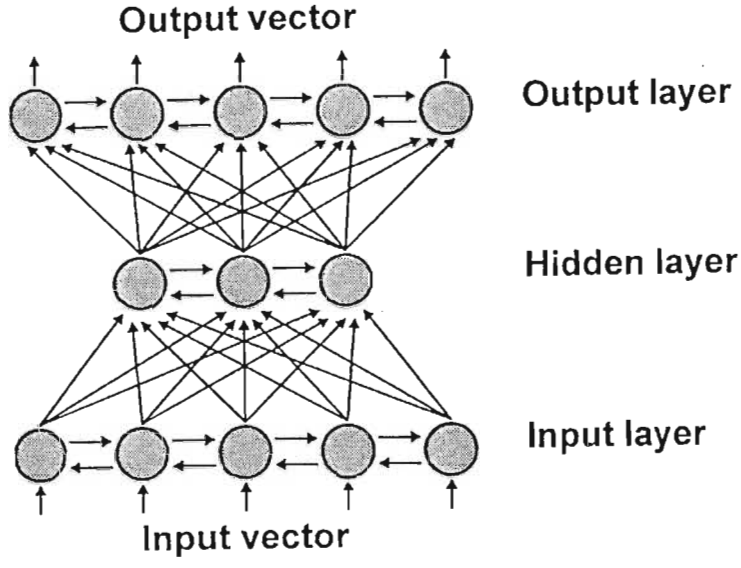


Figure 5.12 Multi-layer co-operative/competitive neural network structure (Maren *et al.*, 1990)

5.2.3 Macrostructure

"Often a single neural network is incapable of solving all aspects of a problem owing to the nature of the data and the problem involved. The solution is found in segmenting the problem into several neural networks that are each apportioned a separate task. This allows the neural networks to act independently yet combine their results either through a neural network or a logical function. In solving large problems, this "panel of experts" approach is often the only way of ensuring cohesive data processing, analysis and decision making. The parameters involved in the choice of macrostructure include, the number of neural networks, the sizes of these networks, the types of inter-neural network connection (feedback, feedforward, lateral etc.), the degree of connectivity and the logical gating or non-linear neural network-based opinion fusion.

"A generic structure of a multi-neural network system is shown in Figure 5.13. When designing such systems of neural networks the emphasis is to highlight the positive contributions of individual neural networks and reduce the effect of their weaknesses. Examples of these networks are the Hamming Network and the Counter-propagation Network" (Maren *et al.*, 1990).

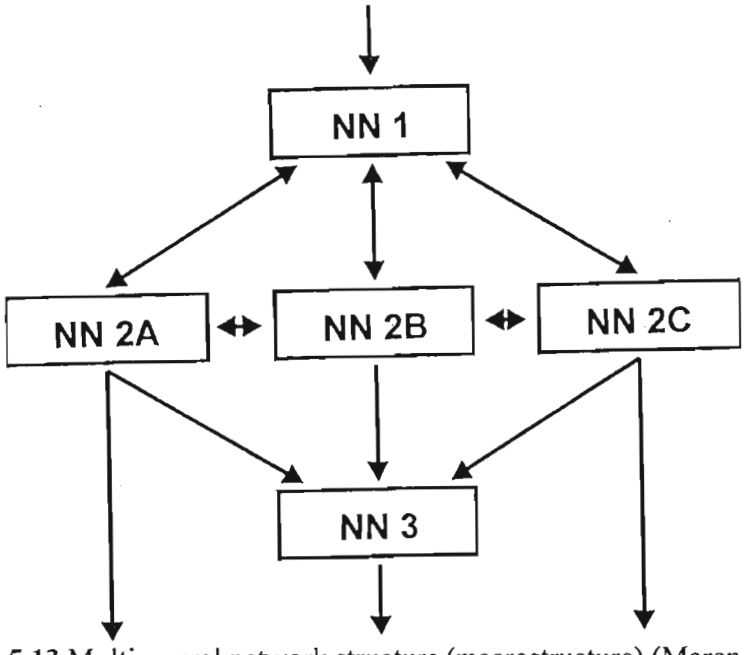


Figure 5.13 Multi-neural network structure (macrostructure) (Maren *et al.*, 1990)

5.2.4 Neural network learning

Learning is an intelligent process. Many computer systems have been programmed to use rules and facts to make conclusions. They use the same rules used by experts to do the job. These systems are called expert systems. Thus their application is limited, although they have a degree of intelligence.

Learning is a unique and important indication of intelligence but it has rarely been achieved through computing. However, simulating the biological model of the brain has developed successful learning programs and this is the field called neural networks.

Learning can be defined as a change in behaviour due to training procedures, where performance is measured before and after training and the difference indicates how

much is learned. Behaviourists refer to learning as conditioning of which there are 2 types:

The first is known as classical conditioning of which Pavlov's experiment is an example. The meat was the unconditioned stimulus with saliva production the unconditioned response, while the bell ringing was the conditioned stimulus with saliva production the conditioned response.

The second is known as operant conditioning and this shall be important in the neural networks used for our problem. Learning is dependent on feedback, referred to as reinforcements. These are stimuli that increase the frequency of a given behaviour with punishment producing the opposite effect. Reinforcement and punishment serve as feedback and are the basis for learning. Shaping, the behavioural training technique that progressively teaches new responses, is accomplished by reinforcing increasingly more approximate behaviours. Note that there is a backwards order of the training procedure.

Learned behaviour does not occur in circumstances isolated to the learning environment e.g.: although the dog may learn to sit in the training lab, he can also perform the sitting trick at the dog show. The situation is different but the dog can sit on the appropriate command. This phenomenon is called generalisation and it is also the ability to respond appropriately when the command or input is noisy. The response is to the message not the irrelevant noise. Also, generalisation enables the correct response to different inputs that have similarities to the training stimuli. Learning is related to a pattern within the total input presentation. There must be some consistencies in the patterns that are the essence of what is learned. These consistencies are the invariants of the pattern.

In the past, background noise had a devastating effect on computerised recognition. Matching patterns or pictures was almost impossible to do unless the pictures were identical.

However, one of the most important features of neural networks is generalisation. Neural networks can be relatively insensitive to noisy input. Extraneous bits on an input pattern are not problematic. Appropriate results are often the outcome even

with slightly different input patterns. As long as the general pattern is present, the network can function suitably thus making it suitable to everyday noisy situations.

The Britannica Dictionary defines learning as:
"the modification of behaviour following upon and induced by interaction with the environment and as a result of experiences leading to the establishment of new patterns of response to external stimuli".

The ability to learn is what differentiates neural networks from other signal processing techniques and statistical methods. Learning in a network of neurons is an extension of the process of adjusting weights of individual neurons to meet a global criterion. The regime used to adjust the individual neuron weights in response to a target is called the learning rule or paradigm.

The process of learning commences with randomisation of the neural network weights to small values around zero according to a Gaussian density function. The learning rule dictates the method of adjusting individual weights to minimise the global error between desired and actual response to input stimuli. In order to learn, training data must be presented to the neural network, so that it may extract the underlying relationship between the input and output vector.

The training data are selected from the training data set randomly without replacement. After several cycles of training data presentation and weight adjustment, target performance criteria are met and the weights stored. These represent the "knowledge" of the neural network. Once training is complete, the neural network may be tested on a separate set of test data with which it has not been presented before.

Performance tests on the novel data reveal whether the neural network has sufficiently abstracted the underlying relationships in the data or whether it has merely learnt the training data set.

There are three classes of learning in neural network, supervised, reinforcement and unsupervised learning:

5.2.4.1 *Supervised learning*

Supervised learning requires a teacher, that is a set of input-output pairs that represent the desired response of the neural network to particular input vectors. This data may be measured, or derived from a human (or rule-based) expert. In the first instance, historical (input) data with its resultant (output) effect are used. For the second case, a human or a rule-based expert system may arbitrate on the correct action to take as a target response to a set of inputs. The first method is most frequently used when dealing with problems when it is known or suspected that there is some deterministic relationship between several inputs and the neural network output.

In many cases, it is not possible to obtain input-output pairs of data, for example when one is unsure how many and how different categories of data should be. Figure 5.14 gives a schematic representation of supervised learning. The most famous supervised learning technique is the back-propagation paradigm.

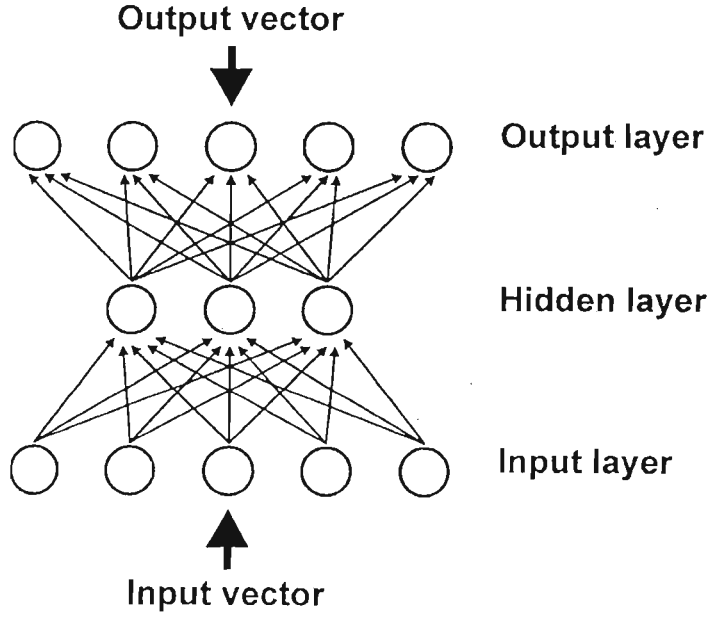


Figure 5.14 Supervised learning paradigm (Maren *et al.*, 1990)

5.2.4.2 *Reinforcement learning*

Although similar to supervised learning, reinforcement learning is much slower. Reinforcement learning (also called Hebbian learning) attempts to emulate the biological method of global reward/punishment in response to inputs. Instead of showing the target response at the output of the neural network to aid learning, this process merely assigns a reward or punishment to the network as a whole. Many training trials are required for the neural network to assign the weights correctly. Compared to the supervised learning case where the learning process is more directed, the reinforcement paradigm is more akin to a directed random search. Figure 5.15 indicates diagrammatically how reinforcement learning works. An example of this technique is the Directed Random Search paradigm and Genetic Algorithm enhanced learning.

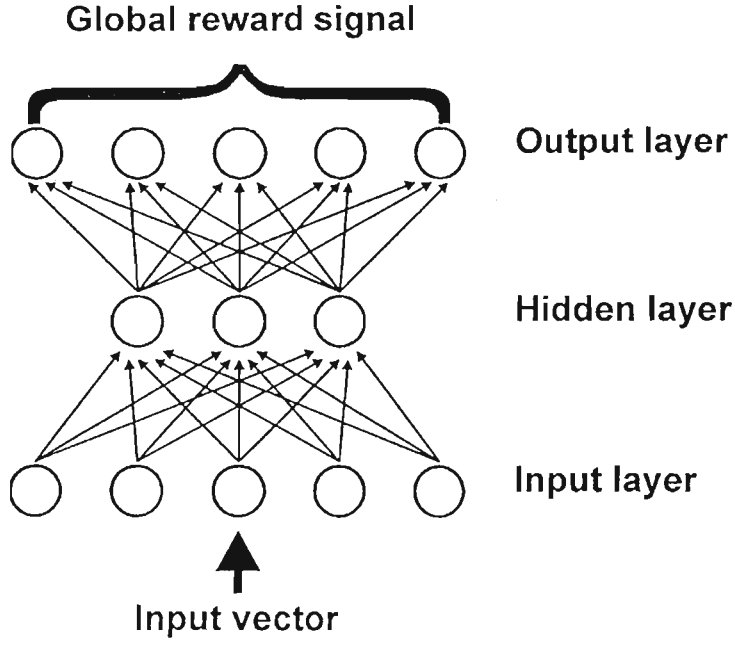


Figure 5.15 Reinforcement learning paradigm (Maren *et al.*, 1990)

5.2.4.3 *Unsupervised learning*

Also known as self-supervised learning, this method allows the data to organise itself into regions of similar characteristics. Thus input vectors that were similar would be projected onto a common classification region, those that were dissimilar onto different classification regions. The two-dimensional plane containing these

classification regions is commonly known as the Kohonen layer after its originator Teuvo Kohonen. Kohonen's neural networks are known as self-organising-maps or simply Kohonen neural networks. A related technique, though not totally unsupervised is the learning vector quantisation paradigm where the "regions" of similarity are uni-dimensional. Figure 5.16 shows a simple self-organising map neural network where similar and dissimilar vectors are matched to separate regions on the Kohonen layer.

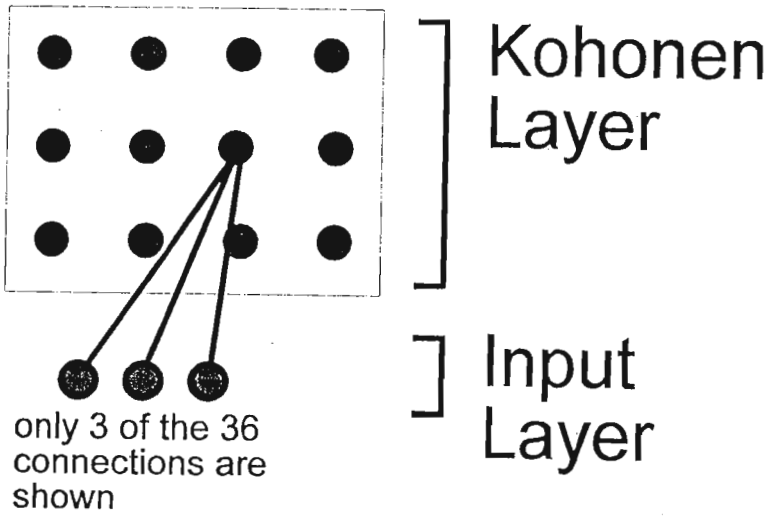


Figure 5.16 Unsupervised (self-supervised) learning paradigm (Maren *et al.*, 1990)

5.3 Back-propagation neural networks

This network has a hidden layer and uses the generalised delta rule for learning. Back-propagation refers to the training method by which the connection weights are adjusted. However, during operation, all information flow is feedforward. A long period of training is required in order to learn the pattern classes.

The key distinguishing characteristic of the back-propagation network is that it forms a mapping from a set of input stimuli to a set of output nodes using features extracted from the input pattern. These mappings can be very complex because the nodes in the hidden layer of the network learn to respond to features found in the input.

Features refer to the correlation of activity among different input nodes, e.g.: suppose that the input layer for a back-propagation network is a 2-dimensional array of binary neurons. Features in this array could mean a pattern of horizontal or vertical node activations, or a corner junction between 2 line segments, or even pairs of activated neurons on opposite sides of the input array. These relationships between the activations of different input neurons provide a basis for a higher-level, more abstract representation of the input information in the hidden layer. Because nodes learn to respond to features as the network is trained with different examples, the network develops the ability to generalise. The ability to make complex distinctions, even when the presented pattern is different from those on which the network was trained, is due to the feature-detection and generalisation abilities which are trained into the hidden layer nodes.

The key-issue is that the hidden-layer nodes must be trained to recognise the right set of features that are sufficiently general so that the network can respond correctly, even when input is different from that which it has previously encountered.

Learning method - the back-propagation network learns to distinguish among different pattern categories simultaneously. Each pattern will have a different type of influence on the change in the connection weights. It is important that the patterns in the training set are presented in such a way that the changes in the connection weights move over time to values which optimise the network's response to all the pattern classes.

The learning rule for the back-propagation network is the generalised delta rule in which the error is used to affect 2 sets of weights (input to hidden and hidden to output). This rule uses the chain rule from differential calculus to calculate the way in which these weights depend on each other.

This is done in 2 stages. Firstly, the connection weights between the hidden or output layers are adjusted. The difference between the actual value of the output neuron and the desired value of the output neuron is used to drive this stage.

In the second stage, the connection weights between the input and hidden layers are adjusted. The basis for this adjustment is different from the one used in the first stage, because neither the system nor the designer knows what the target output of the hidden neurons should be. This is where the back-propagation method offers a unique and effective approach. The back-propagation method uses the adjustments and values to the hidden-to-output weights to help determine the changes made to the input-to-hidden weights.

The goal of the weight adjustments is to reduce the error in the output, or to reduce the difference between the actual and desired output. This error is defined as LMS error. In order to minimise error, differential calculus can be used to obtain a minimum. The training rule can be expressed in terms of a function, delta. This function is applied to every connection weight after each presentation of a training pattern. Repeated use of the delta function will usually lead the network to converge at a set of connection weights which minimises the error for recognising all patterns in the training set.

The performance of the back-propagation network is its ability to generalise. This generalisation and feature extraction is performed by the hidden layer neurons. The knowledge that is stored in the hidden layer is abstracted from the information contained in the input patterns. This abstracted knowledge provides the basis for classifying the pattern into one of the categories available in the output layer. The hidden neurons are actually trained to recognise certain succinct features in the input pattern. Hence the output layer learns to respond to the presence or absence of features in the pattern, and not to the exact spatial or temporal layer of the pattern itself. A wide variety of features can be represented in the hidden layer. The features may be local or they may describe different aspects of the overall pattern.

Back-propagation neural networks are feedforward multi-layer type networks. The back-propagation relates to the method of learning. Back-propagation neural networks rely on supervised training where the training set consists of discrete pairs of input-output vectors. Back-propagation neural networks form a mapping of the input vector to the output vector, the features of this mapping and the relationship

between the input variables is learnt by the hidden layer/s. The layers in a back-propagation neural network are usually fully connected.

The back-propagation paradigm in a multi-layer neural network was co-invented by several researchers in the 1970's and 1980's, (Werbos, 1974, 1988; Rumelhart and McClelland, 1986; Parker, 1985). Back-propagation neural networks are currently the most popular neural network paradigm for real-world applications (Neural Computing, 1993).

5.3.1 Learning in back-propagation neural networks

Back-propagation learns by calculating an error between the desired and the actual neural network output in response to an input. This error is propagated backwards through the network to every neuron. The back-propagated error is used to drive the learning (i.e. weight adaptation) at each neuron.

The rate at which these errors modify the weights is referred to as the learning rate or learning coefficient. Figure 5.17a shows a typical back-propagation neural network structure and Figure 5.17b the notation used for each neuron in the back-propagation neural network analysis (after Neural Computing, 1993.)

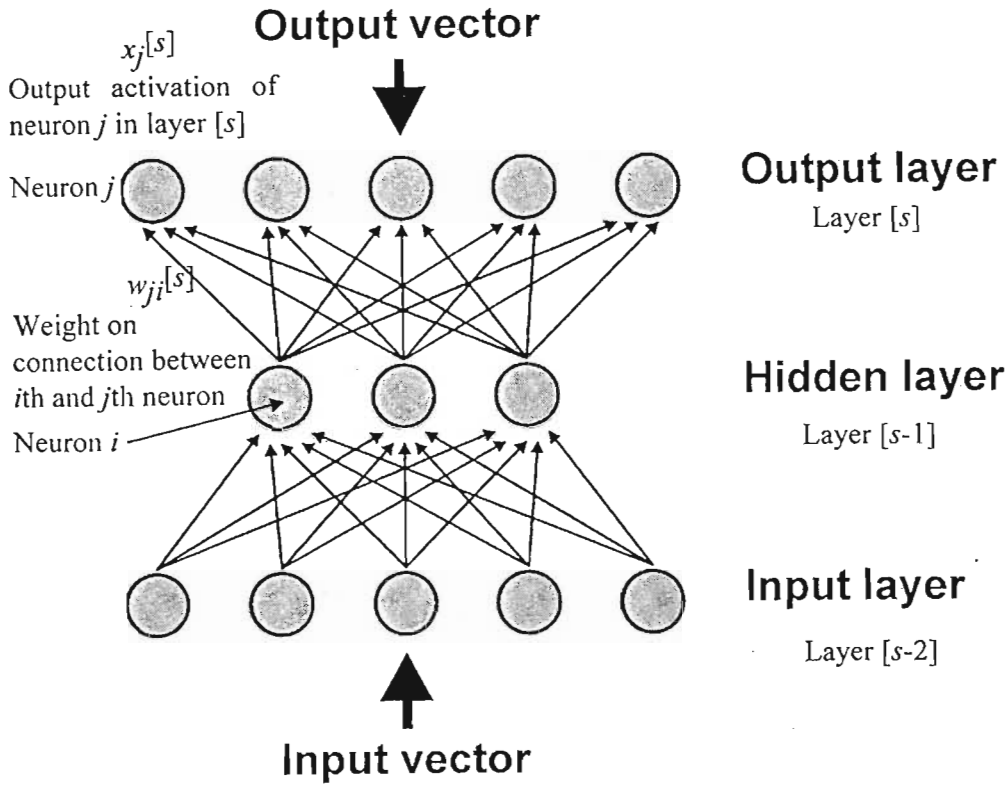


Figure 5.17a Back-propagation nomenclature (Neural Computing, 1993.)

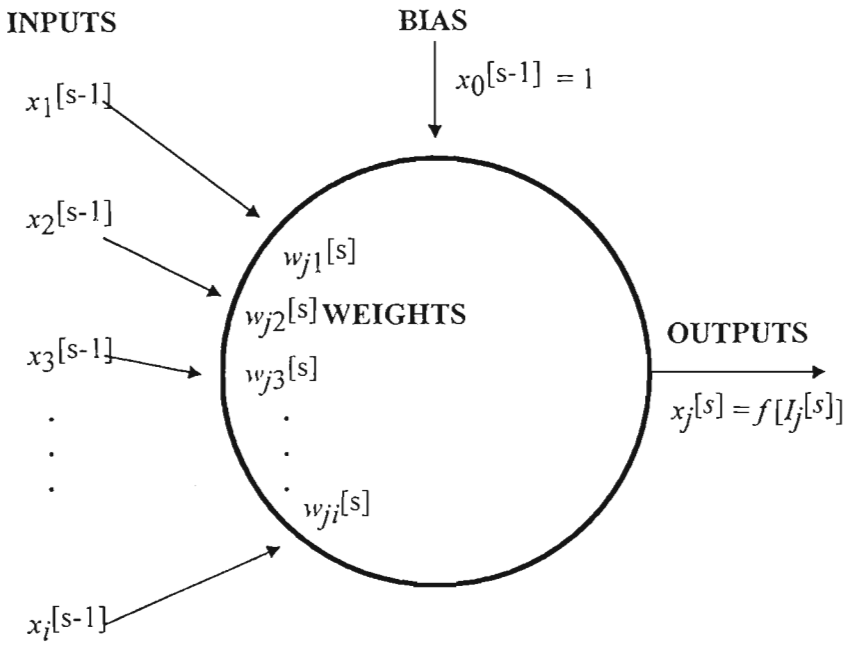


Figure 5.17b Back-propagation neuron nomenclature (Neural Computing, 1993.)

Considering the Figures 5.17a and 5.17b, the output of an individual neuron is given by Werbos, (1988) (after Neural Computing, 1993.):

$$x_j^{[s]} = f\left[\sum_i w_{ji}^{[s]} x_i^{[s-1]}\right] = f\left(I_j^{[s]}\right) \tag{5-3}$$

where

[ ] indicate the layer under consideration

$x_j^{[s]}$ is the output activation of current neuron j in layer [s]

$w_{ji}^{[s]}$ is the weight on the connection between the ith neuron in layer [s-1] to the jth neuron in layer [s]

$x_i^{[s-1]}$ is the incoming activation of neuron i in the previous layer [s-1]

$I_j^{[s]}$ is the weighted sum of the input to the jth neuron in layer [s]

f is the transfer function $f(z) = \left(1 + e^{-z}\right)^{-1}$ for a logistic sigmoid

5.3.2 Back-propagating the local error

The neural network has a global error function E associated with it which is a differentiable function of all the connection weights of the neural network. The parameter that is passed back through the layers is (after Neural Computing, 1993.):

$$e_j^{[s]} = \frac{-\partial}{\partial} \frac{E}{I_j^{[s]}} \tag{5-4}$$

where

$e_j^{[s]}$ is a measure of the local error at neuron j in layer [s]

Using the chain rule twice gives a relationship between the local error at a particular neuron in layer [s] and all the local errors at in the following layer [s+1] (after Neural Computing, 1993.):

$$e_j^{[s]} = f'\left(I_j^{[s]}\right) \cdot \sum_k \left(e_k^{[s+1]} \cdot w_{kj}^{[s+1]}\right) \tag{5-5}$$

Note: this is used in all layers except the output layer (since it has no [s+1] layer).

If the logistic sigmoid transfer function is used in equation (5-5), then the derivative of the logistic sigmoid is a simple function of itself using the basic expansion:

$$f'(z) = f(z) \cdot \left(1 - f(z)\right) = \frac{e^{-z}}{\left[1 + e^{-z}\right]^2} \tag{5-6}$$

If the derivative is near zero (when $f(z)$ is near 0 or 1), then the change in weights is small. These are two stable states. When the derivative is not near zero, there is a corresponding increase in weight change (after Neural Computing, 1993).

If the transfer function is a hyperbolic tangent function, then its derivative may also be expressed with respect to itself using the basic expansion:

$$f'(z) = \left(1 + f(z)\right)\left(1 - f(z)\right) \tag{5-7}$$

Using this equation (5-5) would be modified to (after Neural Computing, 1993):

$$e_j^{[s]} = \left(1 + x_j^{[s]}\right)\left(1 - x_j^{[s]}\right) \cdot \sum_k \left(e_k^{[s+1]} \cdot w_{kj}^{[s+1]}\right) \tag{5-8}$$

Continuing with the logistic sigmoid transfer function, using (5-6), (5-5) can be re-written as follows (provided f is a logistic sigmoid) (after Neural Computing, 1993):

$$e_j^{[s]} = x_j^{[s]}\left(1 - x_j^{[s]}\right) \cdot \sum_k \left(e_k^{[s+1]} \cdot w_{kj}^{[s+1]}\right) \tag{5-9}$$

The summation term in (5-9) is which is used to back-propagate errors is analogous to the summation term in (5-3) which is used to forward propagate the input through the network. Thus the main mechanism is to forward propagate the input and then back-propagate the errors from the output to the input using (5-9) or more generally (5-7) (after Neural Computing, 1993.).

5.3.3 Minimising the global error

To fulfil the aim of minimising the global error E by modifying the weights, use is made of the local error at each neuron.

If there is a set of weights $w_{ij}^{[s]}$, the global error E must be decreased by either incrementing or decrementing the local weights. The global error E may be visualised as a multi-dimensional surface in n-dimensional space. This may be achieved by using a gradient descent method of searching for an optimal minimum. There are several other rules and heuristics which may be used in place of the gradient descent method to improve learning. These will be discussed shortly as improvements to the back-propagation algorithm.

The gradient descent method gives a measure of changing the individual weights in response to the global error. This equation shows that weight changes are made in response to the size and direction of the negative gradient of the error surface (after Neural Computing, 1993.):

$$\Delta w_{ji}^{[s]} = -l_{coef}\left(\frac{\partial}{\partial}\frac{E}{w_{ji}^{[s]}}\right) \qquad (5\text{-}10)$$

where l_{coef} is a learning coefficient

The partial derivatives in (5-10) can be calculated directly from the local error values. Using the fundamental chain rule and (5-4) (Neural Computing, 1993.):

$$\frac{\partial}{\partial}\frac{E}{w_{ji}^{[s]}} = \left(\frac{\partial}{\partial}\frac{E}{I_j^{[s]}}\right)\cdot\left(\frac{\partial}{\partial}\frac{I_j^{[s]}}{w_{ji}^{[s]}}\right) \qquad (5\text{-}11)$$

$$= -e_j^{[s]}x_i^{[s]}$$

Combining (5-10) and (5-11):

$$\Delta w_{ji}^{[s]} = l_{coef}e_j^{[s]}x_i^{[s-1]} \qquad (5\text{-}12)$$

5.3.4 The global error function

The logistic sigmoid function is needed to specify the local errors at the output layer so that they can be propagated backwards. Suppose vector *i* is presented to the input and vector *d* presented to the output as a training data pair. Let *o* denote the actual output produced by the neural network with its current set of weights. Then a measure of the error in achieving the desired output is given by the square of the Euclidean distance between the two vectors (also known as the mean-squared error) (after Neural Computing, 1993.):

$$E = \frac{1}{2}\sum_{k}\left[\left(d_k - o_k\right)^2\right] \tag{5-13}$$

where k indexes the components of $\underline{d}$ and $\underline{o}$

Here the raw local error at each neuron of the output layer is (after Neural Computing, 1993.):

$$e_k^{[o]} = \frac{-\partial}{\partial} \frac{E}{I_k^{[o]}}$$

$$= \frac{-\partial}{\partial} \frac{E}{o_k} \cdot \frac{\partial}{\partial} \frac{o_k}{I_k^{[o]}} \tag{5-14}$$

$$= \left(d_k - o_k\right) f'\left(I_k^{[o]}\right)$$

$$e_k^{[o]} = \left(d_k - o_k\right) x_k^{[o]}\left(1 - x_k^{[o]}\right)$$

Other error functions may be substituted for this standard one (after Neural Computing, 1993.):

$$E = \frac{1}{3}\sum_{k}\left|\left(d_k - o_k\right)^3\right| \tag{5-15}$$

$$E = \frac{1}{4}\sum_{k}\left[\left(d_k - o_k\right)^4\right] \tag{5-16}$$

These give local errors of (after Neural Computing, 1993.):

$$e_k^{[o]} = \left(d_k - o_k\right)^2 f'\left(I_k^{[o]}\right) \tag{5-17}$$

(quadratic error function)

$$e_k^{[o]} = \left(d_k - o_k\right)^3 f'\left(I_k^{[o]}\right) \tag{5-18}$$

(cubic error function)

(5-13) is for the global error for a *particular* (i,d). An overall global error can be defined as the sum of all the specific error functions. Then each time a particular (i,d) is shown, the back-propagation modifies weights to reduce that particular component of the overall error function.

5.3.5 Summary of the standard back-propagation algorithm operation

"Given a input vector $\underline{i}$ and a desired output vector $\underline{d}$ do the following (Neural Computing, 1993):

1. Present $\underline{i}$ randomly to the input and propagate it through the neural network to the output obtaining vector $\underline{o}$

2. As $\underline{i}$ propagates through neural network, it will set all the summed inputs I_j and output states x_j for each neuron

3. For *each* neuron in the output layer, calculate the scaled local error as given in (5-14) and then calculate the delta weight using (5-12)

4. For each layer [s], starting at the layer prior to the output layer and moving to the layer after the input layer, and for each neuron in [s] calculate the scaled local error as given by (5-9) then calculate the delta weight using (5-12)

5. Update all weights in the network by adding the delta weights to the corresponding previous weights."

5.3.6 Enhancements to the back-propagation algorithm

Cumulative weight update

In cumulative (batch) mode, back-propagation neural networks accumulate their weights over an epoch before they are applied to the neural network. The epoch size is a number less than or equal to the number of training vectors. Standard back-propagation uses an epoch size of one, i.e. the weight changes are applied as soon as they are calculated. Cumulative back-propagation may aid learning by "filtering out" minor changes in weights and updating general weight change trends instead.

Momentum

The difficulty with the gradient descent method is the setting of the learning coefficient. This is because changing weights as a linear function of the partial derivative makes the assumption that the error surface is locally linear. While this assumption may hold in general, in regions of high curvature it does not and may lead to divergent behaviour if the learning coefficient is not sufficiently small. The drawback with a small learning coefficient is slow learning, therefore the concept of momentum is added to solve the dichotomy. Momentum simply indicates that if weights are changing in a certain direction, there should be a tendency for them to continue changing in the same direction. Momentum allows for a smaller learning coefficient yet with faster learning.

By modifying the delta weight to include a fraction of the previous historical value of the same delta weight, momentum is added to the process to enable the "hill-descent" to overcome local minima as shown (after Neural Computing, 1993):

$$\Delta w_{ji}^{[s]}(current) = l_{coef} e_j^{[s]} x_i^{[s-1]} + momentum \Delta w_{ji}^{[s]}(previous) \quad (5\text{-}19)$$

Saturation and f' offset

To improve the learning speed of the standard back-propagation algorithm, a small positive offset is added to the derivative of the logistic sigmoid. The reason is that

the incoming weights of a neuron can become so large that the activation values saturate at their limits (either 0 & 1 or -1 & +1). The derivative of these values on the transfer function is very near zero, so training at that neuron effectively stops. By adding the f' offset, this saturation problem is alleviated.

Extended-delta-bar-delta learning heuristic

The extended-delta-bar-delta heuristic was developed by Minai and Williams, (1990) as an extension to the delta-bar-delta heuristic of Jacobs, (1988). The delta-bar-delta rule uses past values of the gradient to infer the local curvature of the error surface. This leads to a learning rule in which every connection has a different learning rate that is automatically calculated. The extended delta-bar-delta heuristic also calculates a momentum term for each connection, thus automating both learning rate and momentum choice. This paradigm is the paradigm of choice for back-propagation neural network development, though others such as Cascade Correlation (Fahlman, 1988) have particular strengths in specific applications.

5.3.7 Back-propagation summary

- It is a general-purpose non-linear regression technique that attempts to minimise global error.

- Any multi-dimensional can theoretically be synthesised by a back-propagation neural network.

- It can provide a very compact distributed representation of complex data sets.

In conclusion, back-propagation neural networks being trained to elucidate structures from C-13 NMR natural product spectra are particularly challenging due to their cross-disciplinary nature. Furthermore, neural networks should be particularly appropriate for natural product recognition because of their generalisation ability.

5.4 Neural Networks in Chemistry

Neural networks have already been established as a tool used to solve chemical problems (Zupan and Gasteiger, 1991). Considerable research has been conducted on infrared spectral analysis (Wythoff *et al.*, 1990; Alam *et al.*, 1994; Ricard *et al.*, 1993; Robb and Munk, 1990; Munk *et al.*, 1991; Fessenden and Gyorgyi, 1991; Tanabe *et al.*, 1992; Weigel and Herges, 1992; Meyer and Wiegelt, 1992; Lerner and Lu, 1993) as well as the classification of mass spectral data (Lohninger and Stancl, 1992; Curry and Rumelhart, 1990), nuclear spectral analysis (Keller *et al.*, 1995; Olmos *et al.*, 1992), and drift correction in pattern classification (Smits *et al.*, 1993).

^{13}C NMR research has encompassed prediction of ^{13}C NMR chemical shifts (Kvasnicka, 1991; Ankers and Jurs, 1992; Kvasnicka *et al.*, 1992(a); Kvasnicka *et al.*, 1992(b); Doucet *et al.*, 1993; Miyashita *et al.*, 1994; Panaye *et al.*, 1994; Svozil *et al.*, 1995; Ivanciuc *et al.*, 1996; Clouser and Jurs, 1996). Kvasnicka *et al.* have used neural networks as a classifier of monosubstituted benzenes with respect to their ^{13}C NMR chemical shifts (4 different positions on the benzene skeleton). Their neural networks were composed of 11 input neurons (descriptors which describe the structure of a given substituent -X), and 4 output neurons (equal to chemical shifts in ipso, ortho, meta and para positions).

Munk *et al.*, (1996), in their development of the computer-based system known as SESAMI (Spectrum Interpretation in Computer-Enhanced Structure Elucidation) have used ^{13}C signal positions in identifying 85 substructural features and Thomson and Meyer, (1989) have used data points in the region 4.0 to 3.5 p.p.m. of ^{1}H NMR spectra to identify 24 different sugar alditols using neural networks.

CCRC-Net (Complex Carbohydrate Research Center Neural Networks) is an automated chemical object recognition system developed at the University of Georgia Complex Carbohydrate Research Center by Drs. Faramarz Valafar and Homayoun Valafar that is available on the Internet (http://www.ccrc.uga.edu/). Its search engine utilises a neural network based pattern matching mechanism. Much work has been done on the identification of xyloglucan oligosaccharides from their ^{1}H-NMR spectra. The 500 MHz ^{1}H-NMR spectra of oligoglycosyl alditols are

generated by borohydride reduction of the oligosaccharide subunits of xyloglucans. The spectral simplification that accompanies borohydride reduction has facilitated the structural analysis of oligomeric xyloglucan subunits, because it transforms each pair of mutarotationally interconverting oligosaccharides to a single oligoglycosyl alditol. At present, only the spectra of these oligoglycosyl alditol derivatives are included in the CCRC-Net database. Thus, the ^{1}H-NMR spectra of reducing xyloglucan oligosaccharides are not currently recognisable.

Keller *et. al.*, (1995) have used neural networks for nuclear spectral analysis for waste handling around the 1450 square kilometres of Southeastern Washington State. The spectra used were alpha particle spectra generated by alpha spectrometers to identify the presence of airborne plutonium in work environments. They found that all the spectral information was found between channels 40 and 239 and thus these 200 channels were reduced to 20 equal sized channels and normalised to the maximum channel count as shown in Figure 5.19.

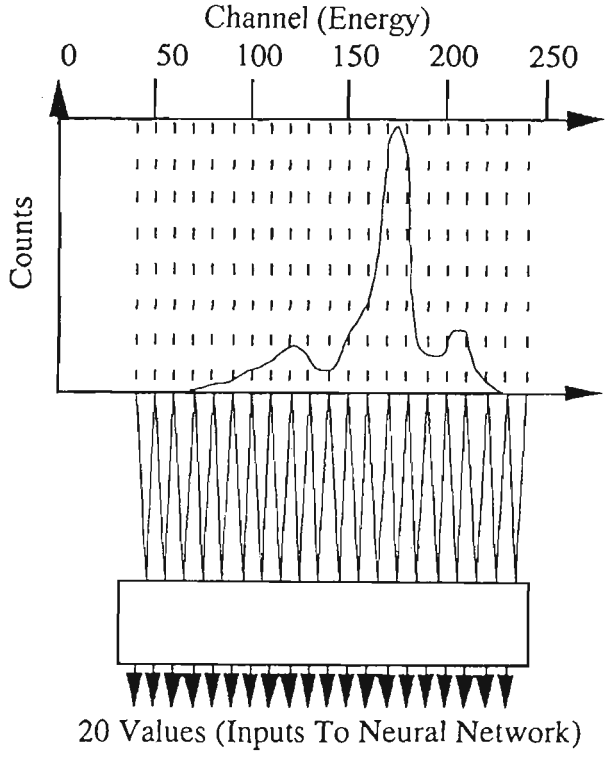


Figure 5.19 Neural networks for spectral analysis, Keller *et. al.*, (1995)

In this research, the intention was to make use of the whole ^{13}C NMR spectrum in the identification process rather than individual peaks or selected regions (i.e. overall pattern), thus making the problem as non-linear as possible. The superiority of humans at visual pattern recognition (i.e.: those of complex natural product ^{13}C NMR spectra) and the link between biological and artificial neural networks provided the inspiration for this research. Back-propagation neural networks have proved to be effective computational tools for pattern recognition and pattern classification. They are particularly well suited to spectral interpretation because the relationship between spectral properties and structural features in a molecule need not be specified, or even known, in advance. The network deduces the relationship during the process of training. This is important, since the rules relating the features of a compound's spectrum to specific structural groupings are often so complicated or so poorly understood that construction of a rule-based interpretation system is impractical (Munk *et al.*, 1996). This work does not reflect an attempt to develop an absolute structural elucidation technique. Neural networks have a remarkable ability to robustly process information containing a degree of uncertainty such as variable, noisy or incomplete inputs often dealt with in natural product chemistry by human experts. It is important to note that the data must not be so thoroughly pre-processed that the problem assumes linearity as other chemometric methods such as discriminant analysis and K-nearest neighbour clustering. Also, no underlying assumptions are made on the behaviour of the data and thus, any variation in the inputs can be accounted for by simply including example input patterns containing such variation during the training process. Another advantage is that most of the intense computation takes place during the training process so that once the neural network is trained, operation is relatively fast and spectra can be rapidly processed without human intervention.

5.5 Example of limonoid and protolimonoid classification using neural networks

Traditionally this has been performed by human experts with advanced analytical and deductive skills and years of exposure to real-world data in the field. In this thesis, it was decided to investigate the technique of classification using these artificial intelligence tools known as neural networks rather than classical classification techniques, for a number of reasons:

- neural networks are excellent pattern recognisers and classifiers
- they do not require NMR spectral data of pure compounds such as those obtained from chemical suppliers
- they base their recognition ability on real data, such as that obtained from an NMR spectrum of a locally isolated natural product extracted in the laboratory
- they are able to cope with the NMR spectra of slightly impure samples
- they are able to cope with peak superposition ie: recognition is not a function of peak position but rather overall pattern perception
- generalisation, unique to neural networks and humans, is essential in tolerating spectral data of natural products
- they are able to cope with the non-quantitative and non-reproducible measurement of line intensities due to the nuclear Overhauser Effect and Fourier Transform pulse intervals
- they are able to tolerate spectra of small amounts of sample which produce weak ^{13}C NMR signals
- they tolerate NMR spectra of compounds which are isolated as inseparable stereoisomeric mixtures or mixtures of esters

There were also other factors that made limonoid, protolimonoid and triterpenoid classification possible in this thesis. Firstly, expert human analysis was available to validate and quantify the performance of the artificial intelligence solution. Secondly, both pure and impure data was available for a wide range of isolated limonoid, protolimonoid and triterpenoid structures. This made system development and validation possible.

5.5.1 Advantages and Disadvantages of neural network classification

Advantages:

1. classify robustly in the presence of noise
2. classify with smooth boundaries between each classification group
3. work robustly in the presence of incomplete, missing or bad data; albeit with lower accuracy
4. work with real data from real processes
5. accommodate wide decision spaces yet retain fine resolution in selected regions
6. make decisions on the input data in parallel
7. map inputs to outputs which are related by highly non-linear relationships
8. operate without *a priori* knowledge of classification groups (unsupervised learning)
9. detect hidden relationships between the input data

Disadvantages:

1. it is difficult, though not impossible, to capture human knowledge explicitly
2. they do not give a deterministic or explicit reason for each decision (the equation *is* the neural network structure itself)
3. they may be skewed or biased by poor data, hence the importance of good and representative data

Chapter 6

Neural Networks Experimental

The purpose of the research presented in this thesis was originally to develop a computer based system using artificial intelligence to recognise and correctly identify the limonoids, protolimonoids and triterpenoids occurring in the species *Turraea floribunda* and *Turraea obtusifolia* of the family Meliaceae. These compounds were isolated during past work by the author, (Akerman, 1991; Akerman *et al.*, 1992; Fraser *et al.*, 1992; Fraser and Mulholland, 1993; Fraser et al., 1994; Fraser and Mulholland, 1995; Fraser et al., 1995) but the research has now been extended to include two new limonoids that have recently been isolated from *T. obtusifolia* seeds by the author, as well as all the limonoids, protolimonoids and triterpenoids isolated from various species within the Meliaceae family by the Natural Products Research team at the University of Natal.

6.1 Data Source

The limonoids found in *T. obtusifolia* and *T. floribunda* that were used in this study (see Appendix I) have wide structural ranges. Some members belong to the highly complex prieurianin group (nymania-1**(I)**, nymania-1 acetate **(II)**, rohitukin-type **(III)**) with rings A and B open (Fraser *et al.*, 1995). The chromatographic details for compounds **I**, **II** and **III** are given in Chapter 3.

Some members belong to the toonafolin group with only ring B open **(IV,V)** (Fraser *et al.*, 1994). Compounds **IV** and **V** were isolated from the seed of *T. floribunda* (57.85g) which were ground and extracted in a Soxhlet apparatus with hexane for 8 hours. Evaporation of the extract yielded a gum (8.58g) which was chromatographed (Merck Art. 7729) using a solvent system of methylene chloride and making it progressively more polar by adding ethyl acetate in 5% successions.

Compound **IV** was obtained as an amorphous material and Compound **V** was obtained as a clear, yellowish gum after a sodium borohydride reduction in order to purify it.

Other members belong to the havanensin group in which all rings of the nucleus remain intact (**VI,VII,VIII**))(Akerman, 1991). Compound **VI** was isolated from the leaves of *T. floribunda* (278.38g). After extraction with refluxing hexane for 8 hours, the extract was evaporated to yield a gum (99.03g). The gum was chromatographed similarly to compounds **IV** and **V** and a slightly impure compound **VI** was obtained as a yellow gum. Compound **VI** was then acetylated using acetic anhydride in pyridine, and after column chromatography using the same method (i.e.: silica gel and solvent system) as that for compounds **IV** and **V**, the major component **VII** and minor component **VIII** were obtained.

The protolimonoids isolated include melianone (present as an anomeric mixture at C-21) (**IX**) and its acetate (**X**) which were isolated from the leaves of *T. obusifolia* (49.27g). The leaves were dried in an oven overnight, liquid nitrogen was poured over them, and the crisp leaves crushed. The resulting fine leaf powder was extracted with refluxing hexane for 8 hours and the hexane extract was evaporated to yield a gum (17.55g). After column chromatography conducted similarly to Compounds **IV** and **V**, compound **IX** was obtained as an orange gum and a minor compound, sapelin F (**XI**) as a clear gum. Compound **IX** (200mg) was then acetylated using acetic anhydride in pyridine to yield compound **X**.

Melianodiol (**XII**), its acetate (**XIII**), melianotriol (**XIV**) and 3α-acetoxy-7α-deacetoxy-glabretal (**XV**) (Akerman, 1991) were all isolated from the wood extract (21.75g) of *T. obtusifolia*. Extraction and column chromatography was conducted as for compounds **IV** and **V** to yield compounds **XII** and **XV** as colourless gums. Compound **XII** (150mg) was acetylated using acetic anhydride in pyridine and after column chromatography, the main product, compound **XIII** was isolated. A sodium borohydride reduction was then carried out on compound **XII** (180.9mg) and after purification of the resulting residue by means of column chromatography, the main product isolated was labelled compound **XV**.

Three other triterpenoids: sitosterol (**XVI**)(isolated from the wood extract of *T. floribunda* (11.25g)), stigmasterol (**XVII**)(isolated from the seed of *T. obtusifolia* (17.02g)) and lupeol (**XVIII**)(isolated from the wood of *Apodytes dimidiata* (18kg)), were also added to the sample set (Akerman, 1991) and are found abundantly in nature.

The rest of the limonoids, protolimonoids, sterols, dammaranes and masticadienoic acids were obtained from other members of the Natural Product Research Group. The limonoids 6α-acetoxyazadirone (**XIX**), 6α-acetoxy-14β,15β-epoxyazadirone (**XX**), 6α-acetoxy-14β,15β-epoxyazadiradione (**XXI**), 14β,15β-epoxyazadirone (**XXII**) and azadirone (**XXIII**), compound **XXIV** and the two protolimonoids compounds **XXV** and **XXVI** were isolated from the bark of *Entandrophragma delevoyi* (Roberts, 1994).

A glabretal (**XXVII**) was isolated from the heartwood of *Aglaia ferruginaea* (Mulholland and Monkhe, 1993).

Two limonoids (**XXVIII**,**XXIX**) were isolated from the seed coats of *Trichilia dregeana* (Iourine, 1996), while a number of limonoids (**XXX**, **XXXI**, **XXXII**, **XXXIII**) (Iourine, 1996) were isolated from the bark, wood and seeds of *Ekebergia capensis*. Two of these limonoids (**XXX**, **XXXI**) were isolated as inseparable mixtures of esters.

Six dammaranes (a class of triterpenoid) (**XXXIV**, **XXXV**, **XXXVI**, **XXXVII**, **XXXVIII**, **XXXIX**) were isolated from *Astrotrichilia asterotricha* (Mulholland and Nair., 1994). Four masticadienoic acids (**XXXX**, **XXXXI**, **XXXXII**, **XXXXIII**) and two dysoxylic acids (**XXXXIV**, **XXXXV**) (classified as glabretal type protolimonoids as they have the glabretal nucleus but the masticadienoic acid side chain) were isolated from the wood and bark of *Dysoxylum pettigrewianum* (Mulholland and Nair, 1994). A further two turraflorin type limonoids (**XXXXVI**,**XXXXVII**) were isolated from the seed of *Turraea floribunda* (Fraser *et al.*, 1993), an obacunol type limonoid (**XXXXVIII**) was isolated from *Dysoxylum muelleri* (Mulholland and Nair, 1995) and a group 4 limonoid with rings B and D opened (**XXXXIX**) was isolated from *Astrotrichilia asterotricha* (Mulholland *et*

al., 1996). Five glabretal-type protolimonoids were isolated from *Dysoxylum muelleri* (Mulholland *et al.*, 1996) (**L, LI, LII, LIII, LIV**).

Finally four unidentified prieurianin type limonoids were isolated from *Aphanamixas polystacha.*

All of these compounds have very complex ^{13}C NMR spectra. Secondly, the protolimonoids, sterols, dammaranes and masticadienoic acids which are all triterpenoids were specifically chosen as they have a similar number of carbon atoms as the limonoids and thus would be most likely to have the closest ^{13}C NMR spectral profiles. The added problem of structural similarity in many instances makes the task of computer-based classification very difficult.

Lastly, flavonoids and coumarins were added to the data set as their ^{13}C NMR spectra have a totally different profile and, being ubiquitous in all plants, are often present in the extracts.

Kaempferol 3-O-glucoside (**LV**) was isolated from the leaves of *Ekebergia capensis* (Monkhe, 1991).

Four novel coumarins were isolated from *Ekebergia pterophylla* (**LVI, LVII, LVIII, LIX**) (Iourine, 1996).

The following flavonoids were added to the data set to boost the flavonoid representation: myricetin 3-O-galactoside (a flavonol) (**LX**), quercetin 3-O-glucoside (a flavonol)(**LXI**), isoaffinetin (a flavone)(**LXII**), luteolin 7-O-glucoside (a flavone glycoside)(**LXIII**), scutellarein 7-O-glucoside (a flavone glycoside)(**LXIV**) and apigenin 7-O-β-D-glucoside (a flavone)(**LXV**) (Heller and Forkmann, 1988).

^{13}C NMR data was prepared from 69 pure and 66 impure compounds isolated from the Meliaceae family. The 69 pure compounds comprised: 28 limonoids, 17 protolimonoids (of which 7 were glabretal-type protolimonoids), 13 triterpenoids (which included plant sterols, dammaranes and masticadienoic acids) and 11 "others" which were the flavonoid/coumarin set. The distribution of classes was

similar in the 66 impure dataset, which ranged from having a few peaks not visible in the spectra of the coumarins, to as many as 70 extra visible peaks in the spectra of some of the other compounds. The dataset for training and validation therefore consisted of 135 records, one for each compound. In those instances where compounds were isolated as stereoisomeric mixtures (a glabretal-type protolimonoid and melianone), all the peaks representing the one stereoisomer were included in the pure compound. In the case of the isolation of some of the glabretal-type protolimonoids as mixtures of cinnamoyl and benzoyl esters in 1:1 proportion and the isolation of some of the ekebergin type limonoids as mixtures of nicotinate and 2-methylbutyrate esters in 1:1 proportion, all isomeric peaks were also included in the pure compound. However, where there was a mixture having 65% cinnamoyl and 35% benzoyl ester proportions, only those peaks representing the predominant isomer were taken to represent the pure compound.

The size of the sample set in this thesis has a direct bearing on the size of the neural networks. Widrow (Widrow, 1960), the inventor of the Adeline neural network, the predecessor of the back-propagation neural network, and Rummelhart and McClelland (Rummelhart & McClelland, 1996) the key co-inventors of the back-propagation neural network developed the theory relating the number of weights in a fully-interconnected neural network to the size of the sample set. The weights represent the unknowns that have to be solved for through the iterative learning process.

The number of weights are defined by:

$$W = \sum_{i=1}^{n-1} (a_i \times a_{i+1}) \qquad (6\text{-}1)$$

Where:

W = number of weights

n = number of layers in the neural network including the input layer, layer 1

a = number of neurons in each layer

The sample set is then used to determine these unknowns. Widrow determined that an approximate relationship exists between the number of weights W and the size of the sample set S. Depending on the complexity of the problem, the ratio of S to W should be between 2 and 10, i.e. there should be 2 to 10 samples for every weight in the neural network.

This leads to two important points, firstly, the size of the neural network should be constrained to be as small as possible, and secondly, the size of the neural network should be checked to ensure adherence to this criterion. This elementary check was conducted as routine. Several steps were taken to ensure that the ratio was adhered to:

1. The input data was preprocessed by binning to reduce dimensionality without reducing information required for classification. In either extreme, including every raw ppm peak would have increased the size of the neural network, an increase that could not be afforded in the size of the sample set. In fact the size of the sample set of impure and pure compounds, not merely literature-derived "clean" compounds, was the basis upon which many of the pre-processing and neural network decisions were predicated.

2. The data at the input to the neural network was checked for sensitivity to the output. For example, if a consistent peak occurred in all compounds at the similar position with similar relative intensity to the other peaks, then it carried no information that would aid discrimination between the compounds, then that input bin was not used. Similarly if there was a random fluctuation in such a peak which carried no definitive information to assist classification, then that region would be unused. This process was carried out in the development phase of the neural networks. This meant that the input size of the neural networks was usually much less than the total number of bins used. This is an obvious source of "saving weights" since many areas are either consistently zero-valued or near zero valued or lack "decision information".

3. The neural network architecture (paradigm) used was not the standard back-propagation algorithm of 10 years maturity. Two significant adaptations were used namely the extended delta-bar-delta learning heuristic derived from the work of Jacobs (Jacobs, 1988), by Minai and Williams (Minai & Williams, 1990), and the neural network enhancement algorithm, Cascade Correlation, developed by Fahlmann (Fahlmann, 1988). The effect of the first was to learn quickly and well requiring the minimum of hidden neurons to achieve satisfactory learning. The second technique builds the optimal neural network

size by commencing with no hidden neurons and then adding neurons are required for training. This is continued until there are diminishing returns for each added neuron. These techniques ensure right sizing of the neural network, i.e. not arbitrarily guessed at, and secondly ensure that the least possible number of neurons (hence weights) are assigned.

4. The neural network can be constrained, if required to less than complete connectivity between each layer of neurons, i.e. a sparse number of connections can be made to save the number of weights. To do this, an analysis of the sensitivity of each weight as a contributor to the output decision is made. The weights with little or no contribution during training are removed. Analysing the weights to each neuron can also perform this process. If the effect of the neuron is negligible, then it may be "pruned" using a procedure developed by Samad (Samad, 1989). The result is that fewer neurons hence fewer weights are required.

These measures ensured that a superior performance of the sample size to weight number ratio was maintained, which allowed perfectly acceptable neural network development and results with a size-constrained data set.

Bearing in mind the excellent performance of the neural networks and the methodology described above, a number of literature-sourced compounds were added to boost the sample size by 10% and provide extra testing data. The data was obtained using NMR instruments ranging from 300-600 MHz.

Two triterpenes, Sapelin A and Entandrolide (Okorie & Taylor, 1977), a triterpenoid Meliavolin (Zeng *et al.*, 1995) and three secomultiflorane-type triterpenoid acids i.e. Bryononic acid, Secobryononic acid and Secoisobryononic acid were taken from literature (Kosela et al., 1995). Five quassinoids, Vilmorinine (Takeya *et al.*, 1997), Ailantinol A, Ailantinol B, Shinjudilactone, and Ailanthone were added from the American Chemical Society's Journal of Natural Products (Kubota *et al.*, 1996). Ten triterpenoids from *Adina rubella* (Fang *et al.*, 1996) were also added as well as Xylocarpin, Ruageanin B and Ruageanin D (Mootoo *et al.*, 1996), three limonoids (tetranortriterpenoids), for completeness. The triterpenoids from Meliaceae were used, but that the quassinoids were not since the purpose of

the thesis was to demonstrate the principle of neural network classification as applied to Meliaceae and not as a general classifier of any compound from any possible source. The data was added to the test data set.

The NMR spectrometers used were: by Zeng *et al.*, 1995 - Varian VXR-500S (500 MHz), by Takeya *et al.*, 1997, - Bruker AM-400 (400 MHz), by Kubota *et al.*, 1996 - Varian VXR-500/JASCO GSX-500/JEOL Alpha-400 (500, 500, 400 MHz, by Fang *et al.*, 1996 - Bruker AMX-600 (600 MHz) and by Mootoo *et al.*, 1996 - Varian Unity 500 (500 MHz).

The results of these additional compounds can be found in **EXPERIMENT A**, in the Appendix.

6.2 Data Set Size

It is correct to state that the neural network ability to generalise is degraded if the data set is not large enough. Neural networks do not "predict" *per se*, in the sense of extrapolation. The training data defines the boundaries in multidimensional space, of a surface on which all the "points" of the training set lie. The purpose of training is to optimise the shape of the surface to best approximate the relative positions of the data in multidimensional space. Once this is correctly done, any new unseen data presented to the neural network is estimated based on the surface delimited by the data set. Thus neural network "prediction" is in fact interpolation, albeit in areas *between* the data points which have not explicitly been defined by data, but are rather estimated. Figure 6.1 illustrates this.

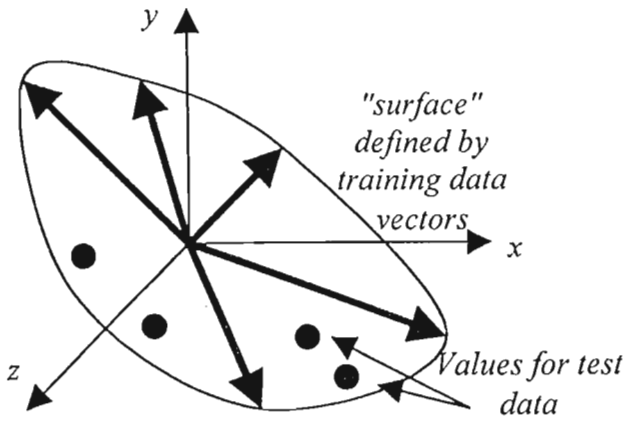


Figure 6.1 Prediction by interpolation of values for test data based on the surface contour defined by the training vectors

Two important deductions can be made as illustrated in Figure 6.1. Firstly if there are insufficient training data points then the detail of the decision surface is poor, i.e. it lacks granularity. Secondly if the training data does not represent both the bounds of the surface as well as points within it, then there will be areas where interpolative decisions are not possible owing to sparse data. Neural networks typically only extrapolate up to one standard deviation from each "point" in multidimensional space defined by the training data vector. So the rule relating the number of weights required (and hence data set size) required to produce a general solution to a problem, depends heavily on the complexity of the decision boundaries and the decision surface to be defined. If they are complex, then more data is required to satisfy the needs of more weights (each weight adds granularity to the decision surface much like increasing the order of the polynomial used in general regression techniques).

In the case of the neural networks used in the thesis, all were classification neural networks. This meant that the importance of a highly defined decision surface was reduced since regions in space defining each grouping were used, rather than absolute values on the decision surface as is the case when the neural networks are used for regression analysis e.g. the estimation of the *value* of a parameter. With a lower importance on the shape of the decision surface comes a lower demand on the number of data points required to define it. But what of the decision boundaries

between the classes? Normally these would be highly complex surfaces, but using a pre-processing technique as discussed in the thesis, the inputs were subjected to several standard non-linear transformations and the optimum transformations selected.

6.3 Data Preparation

Pre-processing of the raw data is a critical step in designing good neural networks as the spectra have to be represented as economically as possible without losing significant information. Robb and Munk, (1990), have shown that neural networks can learn to associate functional groups with peak positions in infrared spectroscopy. A binning technique was employed by Keller *et al.*, (1995) for alpha spectral analysis. Various data pre-processing options exist for spectral data: those methods that transform individual input sample vectors and those that transform wavelength, frequency/intensity input vectors (Alam *et al.*, 1994), but the input vector must be of consistent dimension. However, a minimum of data pre-processing is desirable in order to evaluate the capability and efficiency of the neural network itself. Too much pre-processing could oversimplify the classification task making it a simple linearly separable problem. This is undesirable since the neural network would lose its generalising and interpolative ability when dealing with non-ideal data. Furthermore it would impose the data inferences forced by pre-processing on the neural network rather than allowing the neural network to infer its own relationships.

To avoid this a simple method of pre-processing was chosen which imposed minimal information loss, minimal input data skewing and the use of a fixed vector size while at the same time reducing the dimension of the input decision space (i.e. the size of the input vector I).

6.3.1 Pre-processing method

The data set consisted of a contiguous table of values, relating peak intensity to peak position. Clearly different samples yielded different numbers, amplitudes and positions of peaks, as well as differing levels of background noise.

The input to the neural networks consists of a vector of dimension n, where n is the number of data points for each compound. Neural networks reason on this data based on:

1) the magnitude and

2) the position in the vector.

This means that when one compound has 10 equal-amplitude peaks all below 100 ppm, and another also with 10 equal-amplitude peaks but above 100 ppm, they have to be presented consistently to neural network to allow for correct discrimination. Thus if the input vector dimension was n=10, then no distinction would take place.

A difficulty however arises in using all 220 ppm of the NMR spectra as inputs. If n=220, then each ppm of the spectrum would be represented by one input to the neural network. In this case, the individual contribution of each significant peak, (essential to manual peak-picking techniques), is lost since the neural network treats each input as equally important to learning. This misses out on the principal peaks and over-emphasises the relevance of spurious peaks. Secondly, the greater the number of inputs to a neural network, the greater the number of weights (w) which have to be calculated ($w = (n \times m) + (m \times p)$, where n = the number of inputs, m = the number of hidden neurons, p = the number of output neurons). Typically, the number of vectors required for training increases in approximately linear proportion to the number of input neurons (and hence weights). Clearly this implies that a reduction in input vector dimension is particularly welcomed when considering the limited numbers of limonoid samples from the Meliaceae obtainable in pure and impure form.

If, however, the input is oversimplified and reduced, then the information content relevant to adequate decision making would be lost. As a result, a balance had to

be found between input vector size and classification performance, the stated goal of the thesis.

Since the peaks in a ^{13}C spectrum of a compound are not totally representative because of superposition due to carbons in the same electronic environment and masking by spurious peaks, an alternative pre-processing technique was sought. After investigation and paper search, the binning technique was decided upon. Binning took the form of clustering all peaks in a predefined ppm range (5, 10, or 15) by computation of their mean value, to produce either 44, 22 or 15 equal-width bins, ranging from 0 to 220 ppm, the natural range of the spectra. Binning was significant for the following reasons:

1) it provided a means of reducing the input vector dimension while maintaining sufficient information for decision-making,
2) it gives a means of dynamically trading off these parameters to determine optimal results by varying bin sizes,
3) the input vectors are of consistent size – essential for input to the neural networks,
4) the relative (i.e. signatory) positions of the peaks are preserved and are consistent for each spectrum,
5) all peaks across the range are contributory, not just those in a specific or restricted ppm band.

6.3.2 Data Binning

Binning is a well tried technique in other spectral analysis applications, and lends itself perfectly to the shape-based reasoning of the neural network, rather than the peak-based reasoning of humans and other explicit computational classification techniques. Neural networks are able to handle any data, regardless of its form, content or relevance. Whether they make any sense of the data is a totally different issue. It is analogous to presenting the same information to an individual for classification in 11 languages and expecting a rational assessment - standardisation of presentation is therefore essential, particularly for classification. This leads directly to the two principal reasons for using binning as a data pre-processing technique.

The first as mentioned is to reduce the data size without reducing the information content so as to accommodate a constrained data set. Secondly, binning was used to standardise the input presentation to the neural network to ensure that each spectrum was presented in a similar (note: not identical) fashion to the neural network. "Similar" is used rather than "identical" since binning technique may introduce errors where, for example, there is an offset and a peak falls into a neighbouring bin. However, it must be understood, that as neural networks are pattern recognisers rather than peak pickers, their decisions are based on the overall shape as a signature of the compound and on the relative positions of the peaks. This also means that the absolute information carried in each individual peak is unimportant since the neural network is recognising its interpretation of the spectrum, rather than the raw latent data available in the source information. See Figures 6.2a&b below:

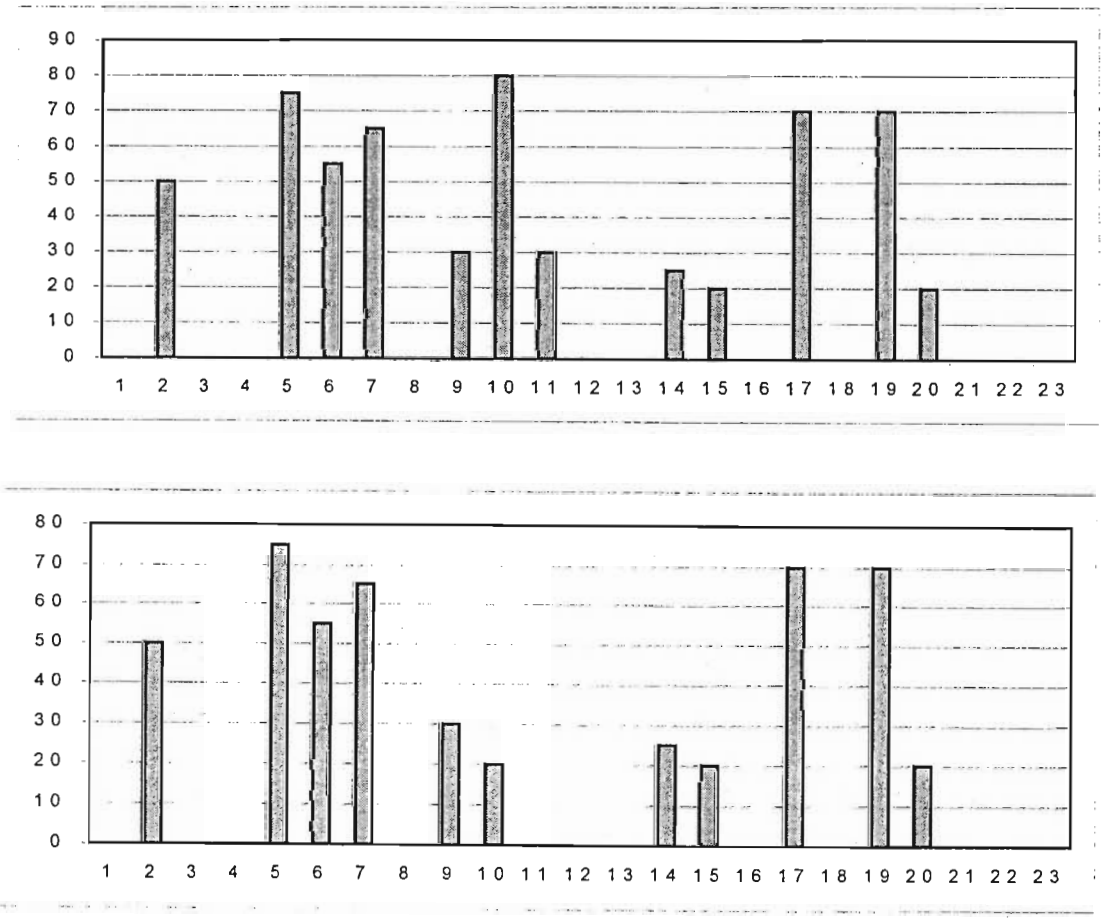


Figure 6.2a&b Comparison between two compounds with identical spectra except two peaks at bin 10,11 (x-axis = ppm bin number, y-axis = amplitude)

A neural network, as well as a trained human would focus on bins 10 and 11 to make a comparison between the compounds since there is no *discriminatory* information in any other areas. Yes, there is information, but it does not make any impact on the distinction between the two spectra. Thus both humans and neural networks do not make their decisions on the raw source data but on the information carried in the perceived differences. In the above, grossly simplified example, a neural network may have learnt that to discriminate between these two, all that is required are two inputs, bin 10 and 11 and the relative (scaled) intensity of bin 10, plus the presence or non-presence of *anything* in bin 11 is the discriminating factor. These "rules" are derived from the data and serve to show that provided there are discriminatory features in the spectra, the neural network will find and use them. These "rules" are based on information only and on no *a priori* knowledge of the source or nature of the data.

This means that provided the reduction in resolution by using binning is not so great as to lose *discriminatory* information, then the effect of binning on classification is zero. So, binning is used as expediency rather than as a necessity.

There are no reasonable practical limitations to using 220 x 1 ppm including computational and complexity limitations. The very real limitation is the size of the data set, which being experimentally derived and then of a specific class of compounds, is naturally limited and requires extra effort to ensure good results on real data. It would have been much easier and more convenient to perform no pre-processing or any of the four-point strategy outlined above and to use a full spectrum. However, the purpose of the thesis is to ascertain the performance of neural networks on real data of clean and dirty (i.e. noisy) nature. This is a much more stringent requirement than the development of a neural network to classify well-elucidated and cleaned journal-derived data - a task that would have taken considerably less time but would have lacked realism and novelty.

The issue then is, how far can one go in reducing the input data size without discarding discriminatory information, and how well can a neural network (or any discriminator) perform classification?

This depends on:

a) how similar the compounds are

b) whether similar compounds are to be isolated or clustered

c) the amount of noise or impurities

d) offsets and

e) bin size

To illustrate the relevance of these effects, a diagrammatical representation is useful.

The inputs to the neural networks consist of vectors of the following form:

For example: 22 bins (10 ppm each)

$$\bar{v} = [x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18} x_{19}, x_{20}, x_{21}, x_{22}]$$

This means that the "space" in which the vector is defined is 22-dimensional. Since this is impossible to represent on paper, a 3-dimensional example (a gross oversimplification) will be used. Consider a vector of three dimensions:

$$\bar{v} = [x, y, z]$$

If each component of the vector represented a "bin" then it could be plotted in 3-dimensional space. Similarly a second vector with different bin-values could be plotted (Figure 6.3):

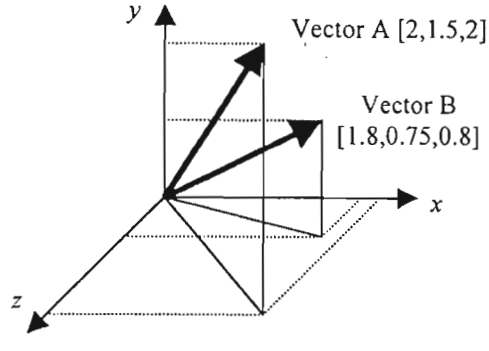


Figure 6.3 Diagram of input vectors in 3-dimensional space for classification analysis

Considering first, **(a)** the level of similarity of the compounds. As already stated, the neural networks do not make their decisions based on all the data in the input vector, but on that data which best discriminates between the vectors. For example if there were a number of vectors, clustered together owing to their similarity as follows (Figure 6.4a) then the neural network would expand or spread the vectors to optimally isolate each member for discrimination (Figure 6.4b):

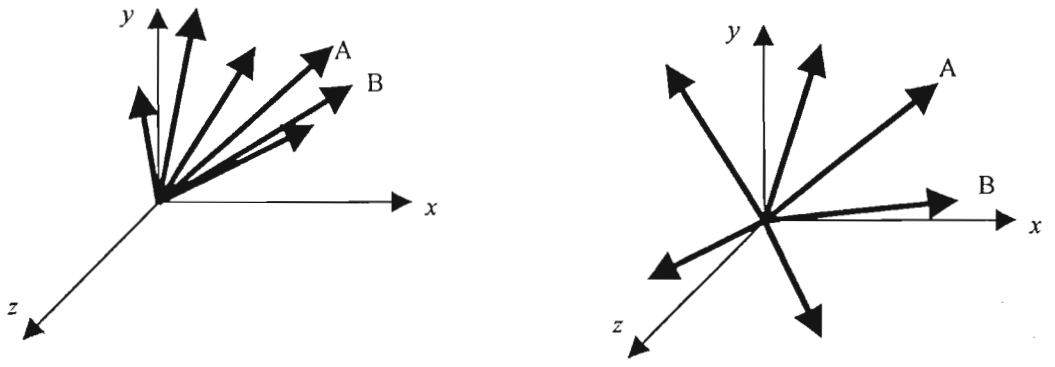


Figure 6.4a&b Neural network expansion of clustered vectors to ease discrimination

This process is simply achieved in the neural network by creating a "sub feature-space" (Figure 6.4b) out of the principal "feature space" (Figure 6.4a), by emphasising only the points of significance of each vector. This normally means that the spread vectors are of lower dimension. In this thesis, the input vectors of dimension 22 or 44 were frequently reduced to vectors of dimension 5 to 7 which represented the number of decision-related inputs required. The process is demonstrated by using the an example of the simplest distance metric for two vectors, namely square of the Euclidean distance:

$$D = \frac{1}{2}\sum_{j}\left(a_j - b_j\right)^2 \qquad (6\text{-}2)$$

Where:

j = vector dimension = (0,1,2...j)

a = vector component of vector A

b = vector component of vector B

Comparing two very similar vectors, A = [0.5, 0.5, 0.1] and B = [0.6, 0.4, 0.1], the first and second components aid discrimination. However, the third component contributes little to the distinction between the 3-dimensional vectors. Reducing the vectors to 2 dimensions and rescaling the values for maximum separation as Figure 6.4b shows a marked improvement in distinction between the vectors. Again this shows that the raw input data has little *direct* significance on the classification capability of the neural network, rather a transformed subset of this data contains the discriminatory information. This condition is considerably different in the case of **(b)** where clustering of vectors is required to provide classification of, for example, limonoids, triterpenoids and others. Here the vectors are not transformed to provide maximum separation. Depending on the proximity of the neighbouring classes greater tolerance of information reduction (Figure 6.5a) can be allowed. Figure 6.5b shows that in this case, the greater the information reduction, the greater effect it will have on the distinction ability of the neural network.

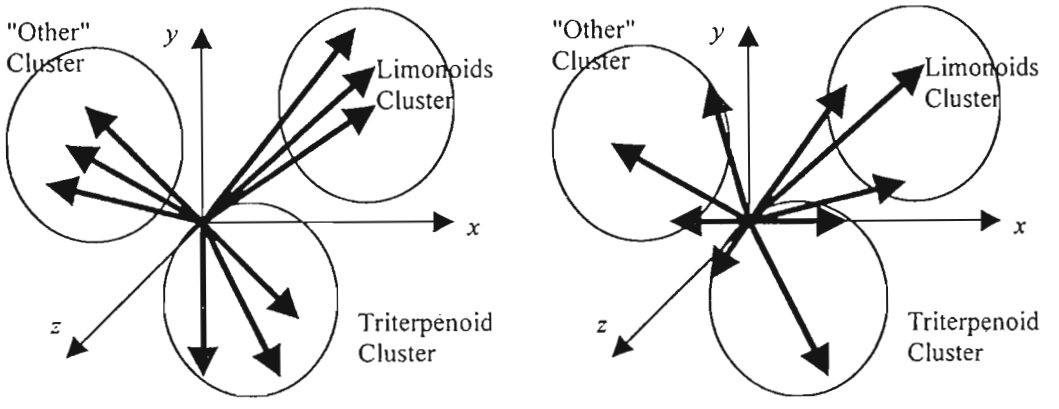


Figure 6.5a&b Neural network clustering of vectors for classification

The third case **(c)**, of the relevance of noise to classification warrants some description. Binning has little effect on spurious peaks originating from impurities, solvent peaks, marker peaks, etc. There are two aspects to "noise" worth considering: random Gaussian (normal) distributed noise arising from the machine and non-Gaussian noise as a result of impurities. These manifest themselves as a low-level random signal evenly distributed across the ppm range in the first case and as isolated "spikes" of much larger magnitude similar to that of the actual compound responses in the second. The data was prepared from tabular printouts from the NMR and not the physical spectral plots. The tabular data was subject to a threshold filter to remove the effects of the instrument noise. Thus the resulting

data was comprised of only large amplitude peaks that corresponded to the compound and to its impurities. Knowing which peaks were related to the compound, the noise peaks could be removed at will. Thus two sets of data were prepared, a noisy set which related to the raw NMR tabular output, and a clean set which was from the "cleaned" spectra.

Two issues arise from this, namely that of threshold setting (related to concentration levels) and magnitudes. As is commonly known, the threshold setting for lower concentrations of compound will have to be higher to isolate the desired peaks than for high-concentrations. The remaining peaks are then scaled by the neural network to fit into a range of 0 to 1 (where 1 corresponds to the value of the largest peak present). Also, in ^{13}C NMR spectroscopy, the magnitudes of the peaks are of relatively small significance as reflected in the literature where only ppm value and peak description is given (s, d, t, q). This means that if the random machine noise is removed, scaling can be performed with any co-efficient of scale without gross-distortion of the spectrum. Thus threshold effects become insignificant in data preparation. Thus the only noise to be considered is the impurity noise. It must be noted that this is non-Gaussian and cannot be "randomly" added to a "clean" compound since they typically related to components of the plant matrix which reflect in very specific areas of the ^{13}C NMR. The inappropriateness of adding random noise is shown in Figure 6.6a and 6.6b. The nature of random noise is that it has a distribution function and not an absolute value. The commonly used root-mean square (RMS) measure of noise is a multiple-instance measure where the distribution of n noise signals are stored and only then is the RMS value calculated.

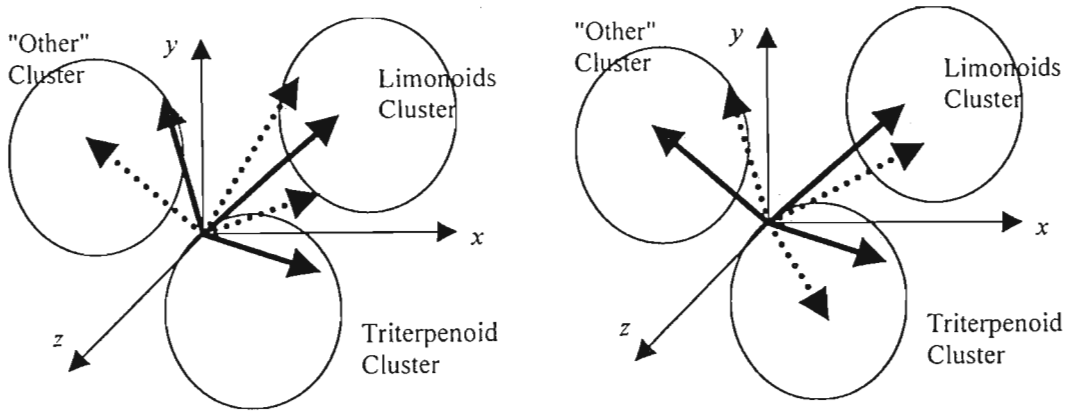


Figure 6.6a&b Effect of random noise on neural network clustering of vectors for classification (solid lines - original vectors, dashed lines - random-noise added vectors)

Referring to Figure 6.6a, the effect of noise was detrimental to the Triterpenoid classification, neutral to the Limonoid classification but actually enhanced the "other" classification. However, this is totally misleading since when noise is again added with a random probability shaped by the Gaussian distribution, the Triterpenoid classification is greatly enhanced, the effect on Limonoids is again neutral but is significantly detrimental to the "other" classification. To eliminate this scientifically, two things have to be done. The effect of the noise has to be stored for a sufficiently large number of instances for a Gaussian distribution with a particular standard deviation to be created for each component of the vector (recall each vector component is a region of the ppm range). Secondly each random process has to be statistically independent for statistically significant results of the noise addition. This means that there must be $n!$ independent combinations, each with a large number of instances to fulfil this criterion. This means that 1.12×10^{21} combinations must be performed for the 22 binned spectra and 2.65×10^{54} for the 44 binned spectra. If the computer could perform 10,000,000 operations per second, this still translates into approximately 1.12×10^{14} seconds of computing time for the former and 2.65×10^{47} seconds for latter (i.e. several million years). Clearly, it would be difficult to perform this scientifically and statistically rigorously.

Therefore, an attempt to create a meaningful measure of the effect of random noise has to be heuristically or experimentally derived. The latter option was chosen since it better reflected the realities of the laboratory-sourced data. The noisy data used to test the neural network performance was therefore the impure spectra containing isolated noise peaks. Measures of this were given in the thesis. Attempts to extend this further by synthetic experimentation are not at all obvious. For example, how does one add truly "random" noise or impurities? Noise or impurities are not random, therefore the added peaks must be subjectively chosen. This is highly sceptical scientifically since it allows the addition of peaks that could be 1) detrimental, 2) neutral or 3) advantageous to classification, i.e. noise of this form is not universally degrading to classification as Figure 6.6a&b showed. So demonstrate this, **EXPERIMENTS B&C**, as found in the Appendix, were performed.

The fourth point **(d)**, to consider is that of offsets. The calibration technique of the NMR can have the effect of introducing an offset in the ppm values of the ^{13}C spectrum. As is stated this may be of the order of 0.1 ppm The effect of this would be to move a peak at say 19.95 ppm from the 10-20 ppm bin into the 20-30 ppm bin. This is possible, however the probability of it happening is extremely small statistically. For example if there are 22 bins each of 10 ppm wide, then there are 23 bin-boundaries where this could occur. If the error is 0.1 ppm then the areas of the spectrum where this could occur amount to 0.1 ppm x 23 = 2.3 ppm Therefore the probability of a peak falling in any of these "indeterminate" zones is 2.3 / 220, i.e. 0.01045. This is insignificantly small. Furthermore, the neural network does not use all peaks to differentiate between compounds, only a selected few that carry discriminatory information, further reducing the probability of the offset peak being required for decision. Secondly the neural network uses a general trend of a number of peaks (in bins) to make a decision; its process is robust and not dependent on the absolute 0.1 ppm accuracy of a particular peak, this is the latent advantage of neural networks over other explicit classification techniques. For that reason, stereo-isomeric mixtures could be used.

The final point to consider **(e)** is that of bin size. As was demonstrated in the thesis, three bin sizes were tried, i.e. 15 ppm, 10 ppm, and 5 ppm As already shown, there is a trade-off between bin size, data set size and loss of information (related to

neural network performance). To demonstrate this compromise, four input vector sizes are used (15ppm = 15 bins (approx.), 10 ppm = 22 bins, 5 ppm = 44 bins, 1 ppm = 220 bins). Consider that there may be say 2 hidden neurons, and three output neurons. Then using the calculation for the number of weights in the neural network (using all inputs, i.e. worst case use of inputs) results in: 36, 50, 94 and 446 weights respectively. This would translate to a requirement of approximately 180, 250, 470 and 2230 samples in the data set. Realistically, the neural networks only selected the relevant inputs to use, so this figure is typically 5 times too large (i.e. 36, 50, 94 and 446 are more likely optimum set sizes). The performance comparison (as given in the thesis) showed that the 15 ppm binned data set performed randomly with comparison to the 10 and 5 ppm data. However the 5 ppm data was consistently better than the 10 ppm data.

The fact that the neural networks' classification performance on the 10 ppm binning was so good indicated that the vectors were sufficiently well clustered to allow information reduction by binning without undue loss of discriminatory features. This is shown in the thesis Table 6.2, where the classification performance enhancement of the 5 ppm bins was only 2-3% over the 10 ppm binned data. This emphasised the point that no smaller than 5 ppm bins were required, or indeed possible with the data set size available. Representing these trade-offs graphically (Figure 6.7a,b,c):

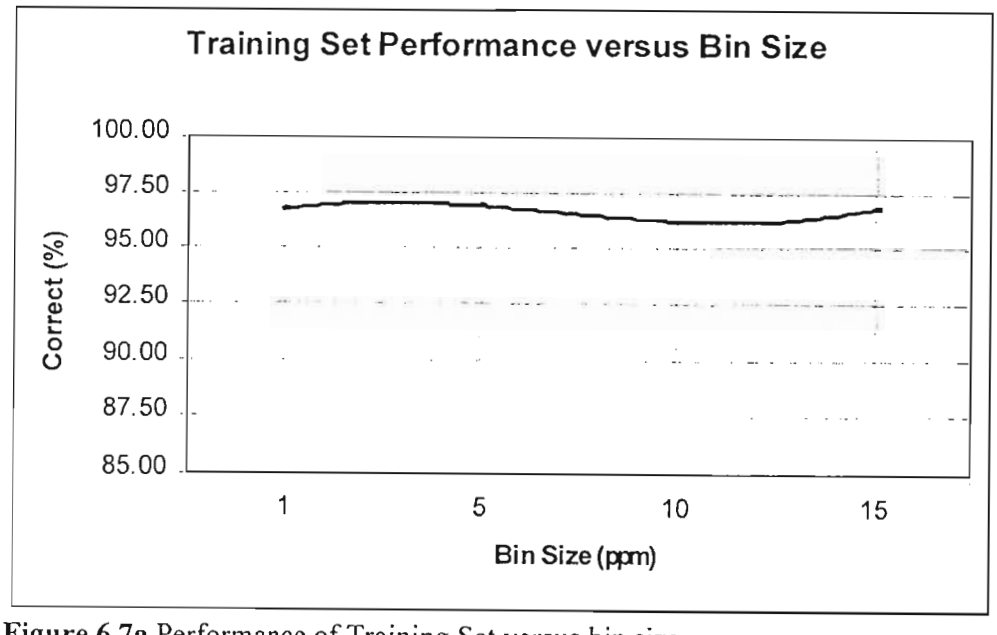


Figure 6.7a Performance of Training Set versus bin size

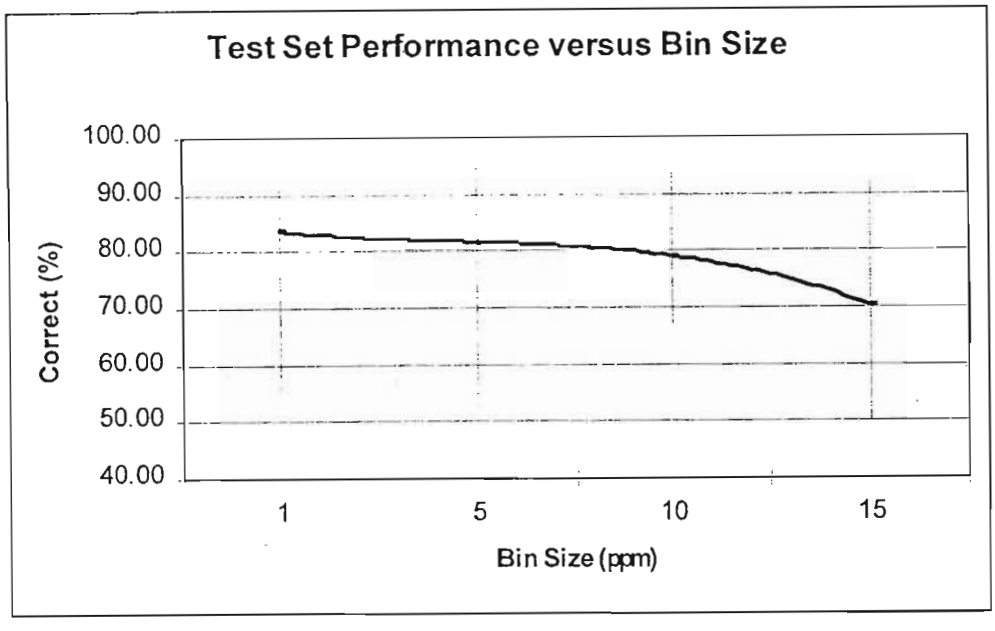


Figure 6.7b Performance of Test Set data versus bin size

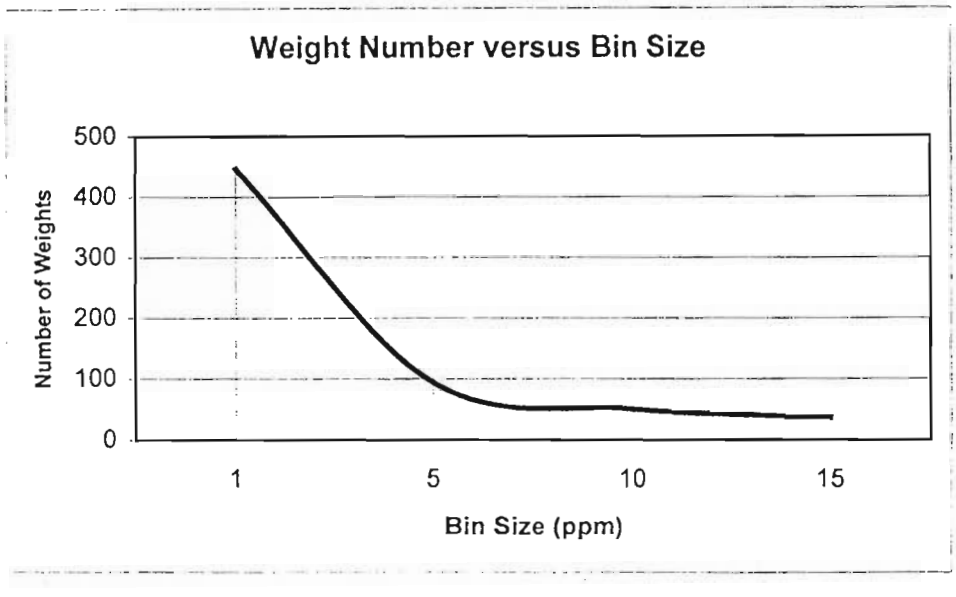


Figure 6.7c Number of neural network weights versus bin size

Figures 6.7a, and 6.7b represent the mean values of the various neural networks' performance for each of the 5, 10 and 15 ppm experiments. The bars at each point represent the spread in performance, i.e. the worst (minimum) and best (maximum) performance of all the 5, 10 and 15 ppm experiments. For both Figure 6.7a,b, values for bin sizes of 1 are derived by fitting a 4^{th} order polynomial curve to the values for 5, 10 and 15 ppm data using Microsoft Excel.

As Figure 6.7a shows, the optimal performance of the training data is very consistent irrespective of bin size. The trend is centred around 96-97% correct. The reason why this should be so is that the neural network algorithm is designed for optimal performance on the data set and training is optimised (i.e. it is not undertrained or overtrained. Figure 6.7a does not reveal any insights into optimal bin size. The small variations around the mean value are due to the fact that each time a neural network is developed, the weights are randomised, and hence do not provide *exactly* the same result each time the neural network is trained, even though all other parameters were kept identical.

Figure 6.7b, shows the trend of diminishing returns for smaller bin sizes very clearly. The performance for 1 ppm binning is 2.5% better than that for 5 ppm binning. Also important to notice is that the performance of 5ppm binning is not significantly better than 10 ppm binning, again underlying the principal that so long as sufficient decision-related data is present, then the effects of binning are negligible. This can be seen plainly in the poor performance of the 15 ppm binned data compared to the 10 ppm data. Clearly the "information" loss in 15 ppm binning was too great for reasonable results. Thus 10 ppm binning should be the upper bound for bin sizes for this data set. What is also of significance is the spread of results. Whereas the mean value of the 5 ppm neural networks was better than the 10 ppm data, the spread in performance was greater.

Figure 6.7a,b seem to indicate that 1 ppm binning may offer better performance, but at what cost? Figure 6.7c indicates the number of neural network weights that would practically be needed to satisfy the input vector sizes of 220, 44, 22, and 14 for 1, 5, 10 and 15 ppm bins respectively. Clearly, based on the available data, 1 ppm binning is impractical, even though it offers marginally better performance. Therefore the use 5-10 ppm bins as an optimal "band" of bin sizes becomes evident. In fact, for the level of performance, spread of results and conservative data set requirements, 10 ppm binning is a good choice.

Naturally as the data set increases in size, so too does the difficulty of classification of more closely separated compounds. Fortunately, as the data set expands, the ability to decrease the size of the bins so as to emphasise finer distinguishing

features improves. There are no practical neural network bounds to the smallest size of the bins except the size of the data set.

A missing peak will not be noticed in a bin where there are already peaks. Fortunately, neural networks do not make their decisions based on the presence or absence of one peak alone, but on the general nature of the data. This holds true so long as the compounds are not 100% identical bar the single missing peak - very few compounds have dissimilarity of one (or for that matter a small number of peaks) alone. They normally differ by a substantial number of peaks. But to demonstrate the argument, consider two hypothetical compounds A & B. Each has been binned into 5 bins each:

A = [0, 2, 3, 2, 2]

B = [0, 2, 0, 2, 1]

If there were say 3 peaks in bin 5 of compound A, and one of the important peaks was missing so that the average dropped from 2 to 1, compound A would still be different to compound B since the discriminatory information would be taken from bin 3 AND 5. If bin 5 yielded no information on which to discriminate, then bin 3 would. This is the robustness of the neural networks (not of binning). The performance of the neural network system and its characteristics should not be confused with the characteristics of binning. It has been emphasised that binning was used for expediency rather than for any real data pre-processing benefit.

The generalising ability of the neural network is completely different from the "generalising" of the binning technique. Throughout the thesis, generalisation is used in the standard accepted form, meaning: the ability to provide a general solution for data with which the neural network has not been trained. By providing "prototypes" of each desired class of compound, the neural network learnt by repeated examination of the data, which features in which bins discriminated one compound class from another. The features would allow an unseen compound possessing a sufficient amount, though not necessarily all, of those distinguishing traits to be presented and still correctly classified. This is generalisation.

What occurred in binning, was not generalisation but rather information reduction. As previously mentioned, it did not matter that the original spectra with 0.1 ppm resolution were compared, so long as enough distinguishing information was still

present after binning. In fact, the effect of binning on the information content of the spectra was minimal since the neural networks reason on the overall shape, and not on individual peaks, an attribute which makes them more noise immune than peak-dependent methods. Binning distorted the overall "shape" of the spectral signature only slightly as is demonstrated in the comparative graphs of 5, 10 and 15 ppm binning (see Figure 6.8, 6.9, 6.10, 6.11). Binning, particularly in 5 ppm widths, still allowed significant peaks however, to contribute to the mean value of the bin.

Furthermore, binning creates few artefacts (spurious products created and not present in original data) but rather smoothes intensity values to the mean, for the range for each bin. This reduces only low amplitude noise and then only to a minor degree. Background spectral noise is typically of a Gaussian (normal) distribution, with a zero mean over the range of ppm values, i.e. white noise. This means that its effect on the mean value of each bin averages out at zero, particularly after presentation of a number of sample vectors to the neural network.

The raw NMR data was pre-processed using the following technique. As nonviscous samples give the sharpest NMR spectra, and the FTNMR instrument maintains frequency accuracy by 'locking' simultaneously to the deuterium NMR frequency, deuteriochloroform ($CDCl_3$) was used to dissolve all samples. Also, these samples are all soluble in $CDCl_3$. Three strong signals in ~ 80ppm region due to deuteriochloroform were consistently removed from all spectra so that during training, the network did not consider them as important peaks as they convey no information about the compounds. The chemical shifts (in ppm) ranged from 0 to 220, so it was decided, to group these into either 44, 22 and in some cases 15 bins, each bin representing a histogram of intensities within that ppm range. It was assumed that the smaller the area of the bin, the better the performance of the neural network. However, in order to justify this assumption and also to evaluate the performance of neural networks with more generalised spectra (i.e.: bigger bin sizes), the neural networks were also trained on the 10 ppm binned data as well as some data which was 15 ppm binned. That is, the intensities for ppm values of 0-5, 0-10 or 0-15 respectively were averaged and placed in bin 44, 22 or 15. Similarly the intensities of ppm values in the range 6-10, 11-20 or 16-30 respectively were averaged and placed in bin 43, 21 or 14 and so on. So the input pattern presented to the neural network consisted of either 44, 22 or 15 numbers representing the 44, 22

or 15 bin-averaged intensity values. Naturally there is a trade-off between bin size and the neural network ability to learn large data spaces, particularly as a function of the number of compounds to be identified and their degree of similarity. Another advantage of this pre-processing technique is that spectra from NMR instruments of higher resolution than the 300 MHz machine used, are presented to the neural networks in a standard uniform format. Below is an example of the unbinned (unprocessed) and 5 ppm and 10 ppm and 15 ppm binned (pre-processed) spectra of pure melianotriol (**XIV**).

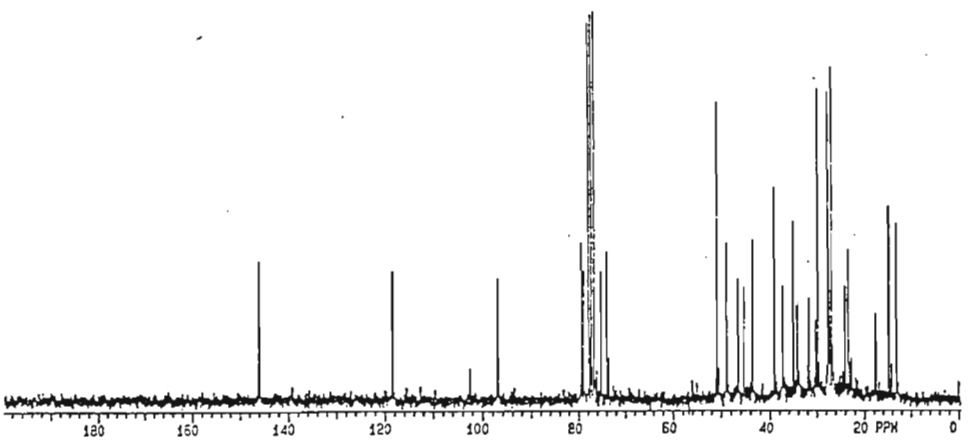


Figure 6.8 ^{13}C NMR spectrum of pure melianotriol (**XIV**)

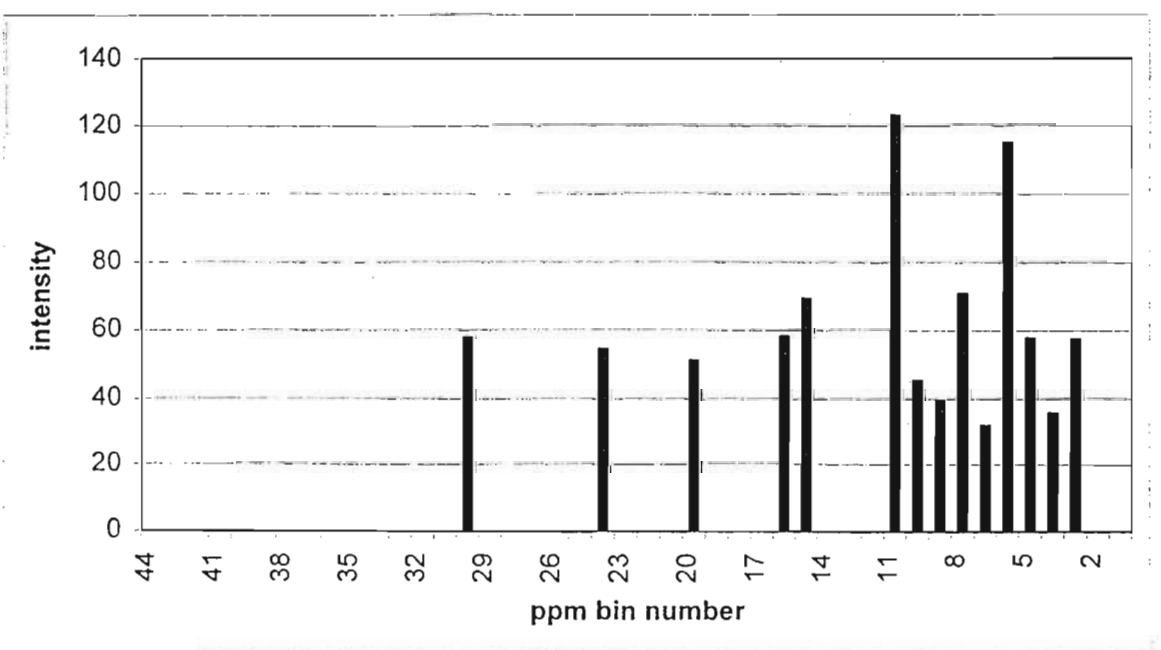


Figure 6.9 5 ppm Binned spectrum of pure melianotriol (**XIV**). The *x*-axis represents the ppm bin number and the *y*-axis represents the intensity.

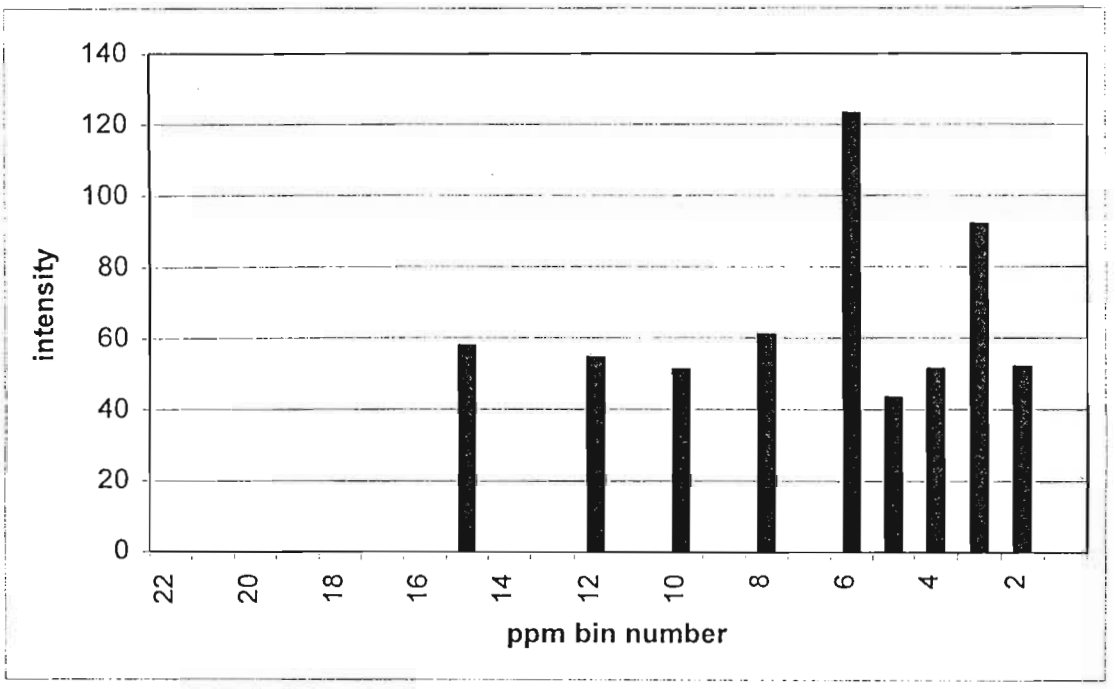


Figure 6.10 10 ppm Binned spectrum of pure melianotriol (**XIV**). The *x*-axis represents the ppm bin number and the *y*-axis represents the intensity.

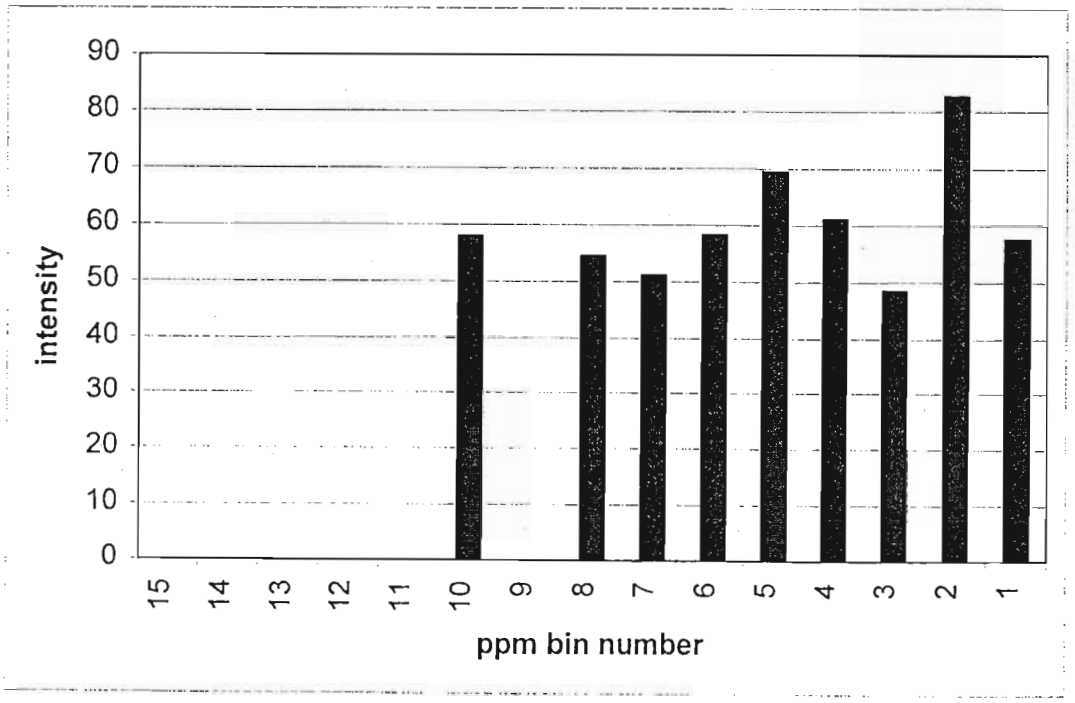


Figure 6.11 15 ppm Binned spectrum of pure melianotriol (**XIV**). The *x*-axis represents the ppm bin number and the *y*-axis represents the intensity.

6.4 Neural network development with *Predict*

A high-level neural network development environment was used to train and test the neural networks. The package called *Predict*, by NeuralWare Inc. USA, was selected as it operates through the Microsoft *Excel* front-end that simplified development considerably by integrating the data manipulation/pre-processing and neural network environments.

The intensities were first normalised to the peak of greatest intensity that was assigned a value of 1. An important factor was that the spectrum could not be too dilute as during normalisation, the noise peaks would become too significant in the overall spectrum, although neural networks cope with this problem better than traditional classification systems. *Predict* uses a variation of the feed-forward neural network architecture, viz. feed-forward with direct connections between input and output as well as cascaded connections (Fahlman, 1988; Bridle, 1990; Kononeko and Bratko, 1991). Many training schemes assume a fixed architecture for the neural network. In other words the number of hidden units is fixed in advance. *Predict* uses a constructive method for determining a suitable number of hidden neurons. This constructive method is referred to as "Cascade Learning". This method is loosely characterised by the following:

1. Hidden neurons are added either one or a few at a time.
2. New hidden neurons have connections from both the input buffer and the previously established hidden neurons.
3. Construction is stopped when performance on an independent test set shows no further improvement.

It is also common for Cascade Learning architectures to have direct connections from the input buffer to the output neurons. As learning rules, two variations of back-propagation are supported, viz., adaptive gradient learning (Ehrlich and Eymard, 1993) and Kalman estimation based learning (Puskorius and Feldkamp, 1991). Since the data that was used to train the neural network was relatively clean, adaptive gradient learning was chosen over Kalman based learning which has its application in estimation of neural network weight values in the presence of noise.

6.4.1 Network Type

Architecture

This showed the number of input, hidden and output processing neurons in the neural net. It was evident that the network structures were fairly small overall, with few if any hidden neurons. This was due to certain transformations which were performed initially on the data to improve their suitability for the neural network learning in *Predict*. This is discussed further under Data Transformations, the result of which is that the problem is made less non-linear. In contrast, the network structures of examples on which ordinary back-propagation neural networks were performed, had a large number of hidden neurons, due to the fact that no pre-processing of the input data had been performed. Thus, *Predict* attempts to simplify the problem as much as possible before the data is entered into the neural network.

Direct Connections

This was selected in all the neural net examples and allowed direct connections from the input layer to the output layer as well as indirectly via the hidden units.

Cascaded Connections

This was also selected in all the neural network examples allowed connections from previously established hidden processing neurons to more recently established hidden processing neurons (Cascaded).

6.4.2 Problem Type

Classification

This was the problem type used as all the outputs represented a finite number of exclusive categories. The output data formed a one-of-*n* code.

Example **{(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)}**

The output data satisfied the following two conditions for each record:

- The field values summed to 1.0.
- Each field value lay between 0 and 1, inclusively.

The one-of-*n* code output data is termed probabilistic.

6.4.3 Data Conditioning

Clean Data

This option was used for most of the neural nets as the data was consistent.

Moderately Noisy Data

This option was used in some instances where it was necessary to generalise the input data more so as to obtain better results.

6.4.4 Data analysis, transformation and variable selection

In the development of neural network models, it is advantageous to analyse and transform the input data even though it has been pre-processed into bins. This helps linearise the input space of the neural network that makes classification in the neural network easier. Furthermore, selecting which transformed variables should be used ensures that data that is irrelevant to classification is not presented to the

neural network. In *Predict*, the data analysis, transformations and variable selection are highly automated. All input data is carefully analysed for bad and missing data, as well as data ranges, distribution and skewness (Plutowski and White, 1992). Thereafter, a suite of transformations (Chatterjee, 1991; Tukey, 1977; Klimasauskas, 1992) is applied to the input data, in this case the mean-intensity ppm bins. These transformations do not add any artefacts to the data but merely separate the underlying relationships in the data into an explicit and a non-explicit component. These transformations help linearise the input "space" (*n*-dimensional) to the neural network. A linearised input space allows for easier and crisper classification of categories. The added advantage of this technique, is that since the input space is more linear than without it, the decision boundaries in *n*-dimensional space can be defined as hyper-planes rather than hyper-surfaces. Hyper-planes can be imagined as cutting planes separating two regions. Mathematically, each hidden neuron in the neural network acts as a point defining this cutting plane. If the plane approximates a linear decision surface as a result of the transformations, then fewer hidden neurons are required to plot the surface. Fewer hidden neurons mean fewer neural network weights, which in turn mean fewer requirements for an extremely large data set. This enabled the unique and limited sample set of limonoids from Meliaceae to be used.

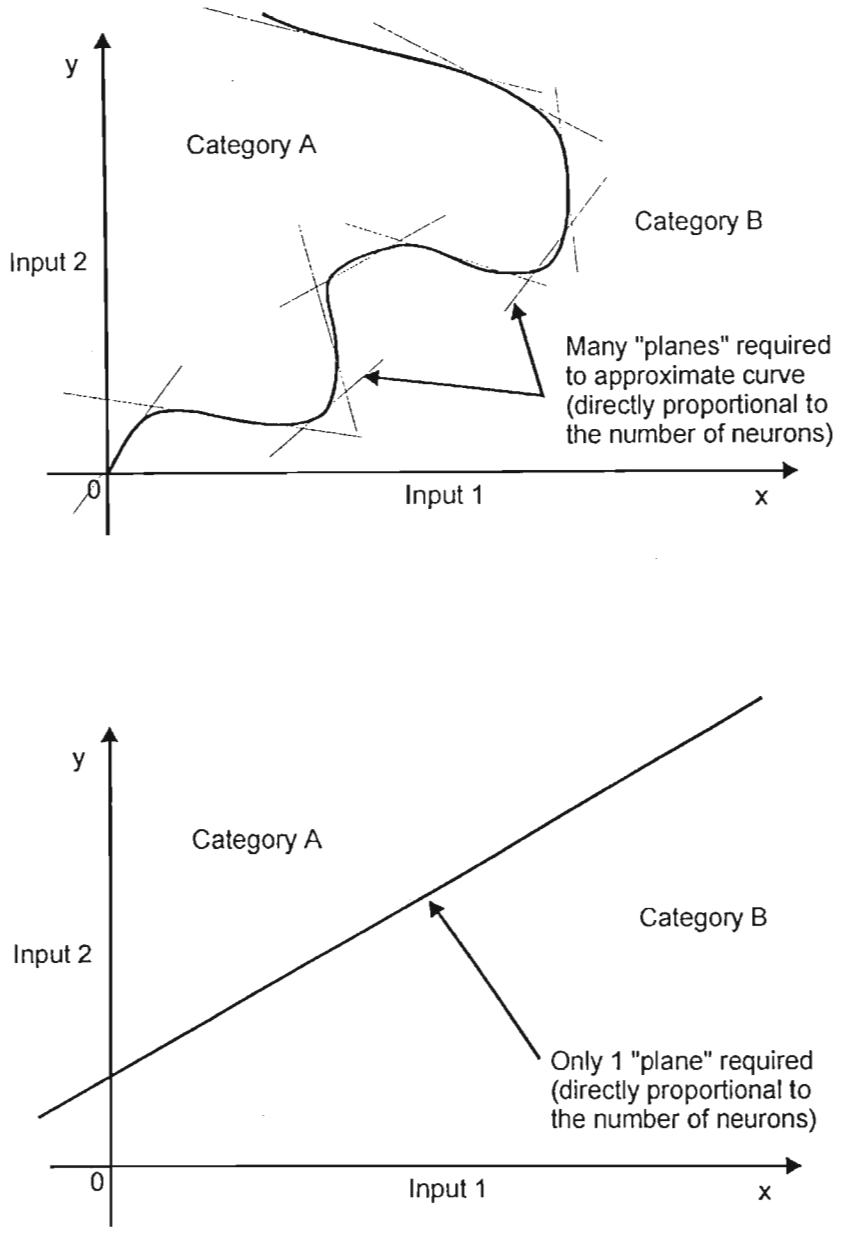


Figure 6.12 Complex and simple (linear) decision boundaries

The inputs, having been transformed, increase the input vector size by the number of transformations performed. Thus a 44 dimensional vector becomes 44 x number of transformations in size. Clearly, not all transformations are appropriate or needed. A non-exhaustive (though close to optimal) search is performed of the combinations of transformed variables to determine which have the greatest relevance or sensitivity on the performance of the neural network. A genetic algorithm (Koza, 1993) is used to perform this function. A genetic algorithm uses evolutionary hypotheses to test and migrate towards an optimum solution in a reasonable period of time. Obviously the combinations of (44 x number of

transformations) is a factorial-type non polynomial-complete problem which becomes computationally intractable to solve using exhaustive searches of all combinations of transformed variables. (eg: travelling salesman problem). The result of this process is a new input vector which typically is smaller than *n*, since the genetic algorithm ignores bins of low or no contribution (e.g. values of zero, or values that are constant for all types of compounds and are hence non-discriminatory). The final "transformation" of the data is a simple scaling of the inputs to +/-1 so as to match the input transfer function limits of the neurons. Transformation increases the dimension of the input space considerably, since each new transformation increases the input dimension by one. Use is made of genetic programming to select which of the transformed variables has the most influence on the neural network classification accuracy (Koza, 1993).

Higher analyses level works harder to find good transformations, and may create more transformations per field. The higher analysis levels also dealt with outliers by creating fuzzy transformations that acted as outlier detectors.

The Data Analysis and Transformation levels were moderate data transformation and comprehensive data transformation which are briefly described below:

Moderate Data Transformation

Using this level, two or three transformed fields were usually generated for each raw continuous field in the data set.

Comprehensive Data Transformation

This level generates a larger and richer set of transforms than the moderate mode and was most often the preferred choice.

6.4.5 Input variable selection

Variable Selection Model was Multiple Regression as described below:

A genetic algorithm was used to find good subsets of the full set of input variables created by the data analysis and transformation component of *Predict*. Genetic algorithms need a fitness function to evaluate individuals in the population. In the case of input variable selection, the individuals are subsets of the total variable set. The fitness of this subset was determined by the variable selection model that was used. The model which was used exclusively was:

Multiple Regression

This option is also known as linear variable selection. It used a logistic multiple regression function to select the model's input variables. It also used a cross-validation procedure to avoid over-dependence on a test set and, lastly, as in the case of all classification problems, a SoftMax output function was used.

Variable Set Size

The genetic algorithm that performs input variable selection looks for sets of variables that act in a synergistic manner as good predictors of the output data. It begins with a population of random variable sets of limited size. As the algorithm progresses, the size of these variable sets will tend to increase if the problem requires larger variable sets.

The different Input Variable Selection levels used were comprehensive variable selection and exhaustive variable selection as stated below:

Comprehensive Variable Selection

This was the variable selection mode used in most cases.

Exhaustive Variable Selection

This mode took substantially longer than the lesser mode and was only used in a few cases. The reason it takes so long is that it builds more complex models with each iteration of the genetic algorithm.

6.4.6 Neural Network Search Level

The different neural network search levels used were:

Comprehensive Network Search

This mode was used almost exclusively as it had a quicker processing time than the Exhaustive Network Search (2 minutes as opposed to 10 minutes).

Exhaustive Network Search

This mode searched harder and longer than "comprehensive network search" and also trained several networks and chose the best one. This mode was used occasionally when the accuracy of using the comprehensive network search was less than 85%.

6.4.7 Data Transformations used were Continuous Transformations, Logical Transformations and Fuzzy Transformations as described below:

Certain transformations were performed initially on the data to improve their suitability for the neural network learning in *Predict*.

The essence of these transformations was to extract the obvious inter-relationships between variables (e.g. if there is a logarithmic relationship in a variable, calculate it first to reduce the learning load on the neural network). This information was added as extra features in the input space. It effectively increased the dimensionality of the problem, but made the problem less non-linear, hence requiring fewer hidden neurons and a smaller neural network structure overall.

Total Fields - The number of input and output data fields in the model.

Active Fields - The number of input and output fields with one or more active transformations.

Total Transformations - The number of input and output transformations.

Active Transformations - The number of active input and output transformations.

Input Transformations used:

Continuous Transformations

The general form of a continuous transformation is:

$$y = s_o f\left(s_i x + o_i\right) + o_o \tag{6-3}$$

where:

f is a continuous function

s_I, o_I implement an inner scaling of the raw data to map it to an optimal sub-domain of f.

s_O, o_O implement an outer scaling so that y lies within a suitable range for the neural net.

In the experimental tables, each Transform was identified by its continuous function f that included the following:

1. Linear Identity function
2. Log - Natural logarithm function
3. LogLog - Log of Log
4. Pwr2 - Square function
5. Pwr4 - Fourth Power function
6. Rt2 - Square root function
7. Rt4 - Fourth root function
8. Inv - Inverse function ($1/x$)
9. InvPwr2 - 1.0 / (Square function)
10. InvPwr4 -1.0 / (Fourth Power function)
11. InvRt2 - 1.0 / (Square root function)
12. InvRt4 - 1.0 / (Fourth root function)
13. Tanh - Hyperbolic tangent function
14. ln $x/(1-x)$ - Log $(x/(1-x))$

Logical Transformations

These are defined as follows:

$$y = T_{max} \quad if\ x > \frac{I_{min} + I_{max}}{2}$$
$$y = T_{min} \quad otherwise \qquad (6\text{-}4)$$

I_{min} and I_{max} are the effective minimum and maximum of the raw data for the field. There may be data points *Predict* considers to be outliers which lie outside I_{min} and I_{max}. T_{min} and T_{max} are the range of the "y" output when the continuous transformation formula has been applied to the I_{min} and I_{max}.

Fuzzy Transformations

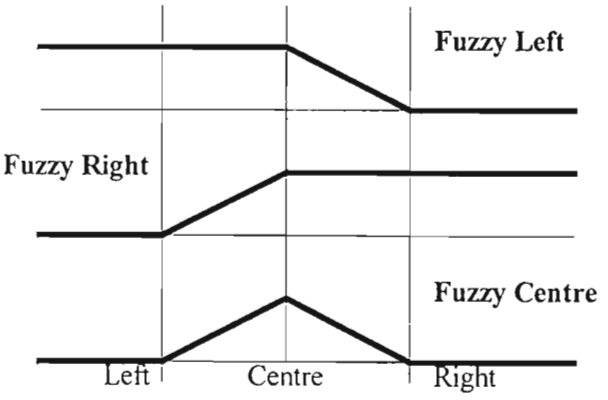


Figure 6.13 Fuzzy Membership Sets

Output Transformations used:

Binary one-of-*n* encoding

6.4.8 Neurodynamics

This refers to the parameters chosen for the performance of the neural network itself.

Available Transfer Functions

The available transfer functions were:

1. Linear.
2. Sigmoid.
3. Tanh.
4. Gaussian.
5. Sine.

The Tanh transfer function was used exclusively as shown in the tables.

Output Layer of the neural network

There were three options for the output transfer function of the neural net.

1. Linear
2. Sigmoid
3. SoftMax

The SoftMax function can only be used for problems where the transformed output data is probabilistic. In other words, for each output record, the sum of values is equal to 1 and each individual output value is between 0 and 1. The SoftMax function was used in conjunction with a relative entropy objective function.

6.4.9 Learning Rule

The learning rule that was used works within the framework of a constructive cascaded architecture:

Adaptive Gradient

The adaptive gradient learning rule uses back-propagated gradient information to guide an iterative line search algorithm. This is a very general learning rule and was used for all the feedforward *Predict* networks.

Weight Decay

This method was used for the "adaptive gradient" learning rule. The weights in a neural net form a highly condensed representation of the training data. During training, the learning rule modifies the weights in response to the training data. If left unchecked, the weights for a processing neuron can latch onto spurious information in the training data, such as data that does not represent a general trend in the input data. By slowly decaying the weights during the course of the training, only the general trends remain encoded in the weights.

Another way of understanding the effect of weight decay is to note that a neural net is usually built with Tanh function building blocks. Large weights exercise the non-linearities in these building blocks much more than small weights. In other words, the more weight decay there is, the closer the model is to being linear.

Values for the weight decay parameters should be kept small (typically less than 0.05). Otherwise any useful information in the data is completely washed out. In practice, weight decay is usually more effective if the weight decay for the output layer is an order of magnitude less than the weight decay for the hidden layer. So, for example, one should couple a value of 0.01 for the hidden weight decay with a value of 0.001 for the output weight decay.

6.4.10 Evaluation

The evaluation technique used during training to evaluate the performance of the feedforward networks was Average Classification Rate.

Classification Rate measures the fraction of correct classifications over all classes and Average Classification Rate measures the average of the class dependent classification rates.

6.4 Learning Vector Quantisation neural network development

Unlike *Predict*, where the network was taught to recognise an exemplar pattern via the back-propagation algorithm, LVQ organises its own representation of categories from the input data. In the LVQ network, a set of vectors is distributed across a space so that the space mimics the probability distribution of the dataset. Thus, the pattern is not spread across a set of neurons but rather each neuron becomes an exemplar for a class (Maren *et al.*, 1990). This method is frequently superior when classifying individuals from a dataset. *NeuralWorks Professional II/PLUS* from NeuralWare Inc. was used to design and train the LVQ networks.

"The learning vector quantisation neural network was originally suggested by Kohonen, (1988) to assign vectors to one of several classes. A learning vector quantisation network contains a Kohonen layer which learns and performs the classification. Learning vector quantisation provides equal numbers of neurons for each class in the Kohonen layer. The basic learning vector quantisation neural network trains and then uses the Kohonen layer as follows (Neural Computing, 1993.):

- In the training mode, the distance of a training vector to each neuron is computed and the nearest neuron is declared to be the winner.

- If the winning neuron is in the class of the training vector, it is moved toward the training vector.

- If the winning neuron is not in the class of the training vector, it is moved away from the training vector (called repulsion).

- During this training process, the neurons assigned to a class, migrate to the region associated with their class.

- In the classification mode, the distance of an input vector to each neuron is computed and again, the nearest neuron is declared to be the winner. The input vector is then assigned to the class of that neuron.

"The basic learning vector quantisation neural network suffers from several shortcomings and variants have been developed to overcome them. These include the addition of a "conscience" mechanism to restrict operation of neurons which win too often and prevent others from participating, a technique to refine the boundaries between classes, a method of including the Bayesian likelihood function and the elimination of the repulsion mechanism. Combined these form a very workable implementation of learning vector quantisation." More details of learning vector quantisation neural network are found in Neural Computing (1993).

6.5 Back propagation neural network development environment

The neural network development environment used was produced by *NeuralWare Inc.* of Pittsburgh USA. The environment which allows cross-platform compatibility between IBM-type personal computers and workstations is called *NeuralWorks Professional II+*. This package supports development of 28 major neural network paradigms with several variations on each. The back-propagation and learning vector quantisation algorithms required for this thesis, were well supported. *NeuralWorks Professional II+* allows building, training, refining and deployment of neural network's. Several features make it particularly useful in development, viz.:

InstaNet - a menu system to select neural network type and parameters

FlashCode - ANSI standard C code generation of neural network for deployment

SaveBest - facility to prevent overtraining a neural network by regularly saving a neural network (i.e. memorisation of training data causing loss of generalisation ability with test data).

Prune - facility during training to eliminate neurons that do not contribute to learning

ExplainNet - feature to explain why a neural network made its decisions and which inputs are important

In addition diagnostic tools such as classification rate, RMS error, weight-histogram etc. help in the development phase. All the features of the development environment are available from an intuitive graphical user interface. This eliminates the need to hard-code neural network paradigms and training schedules. An example of the graphical user interface is shown in Figure 6.14

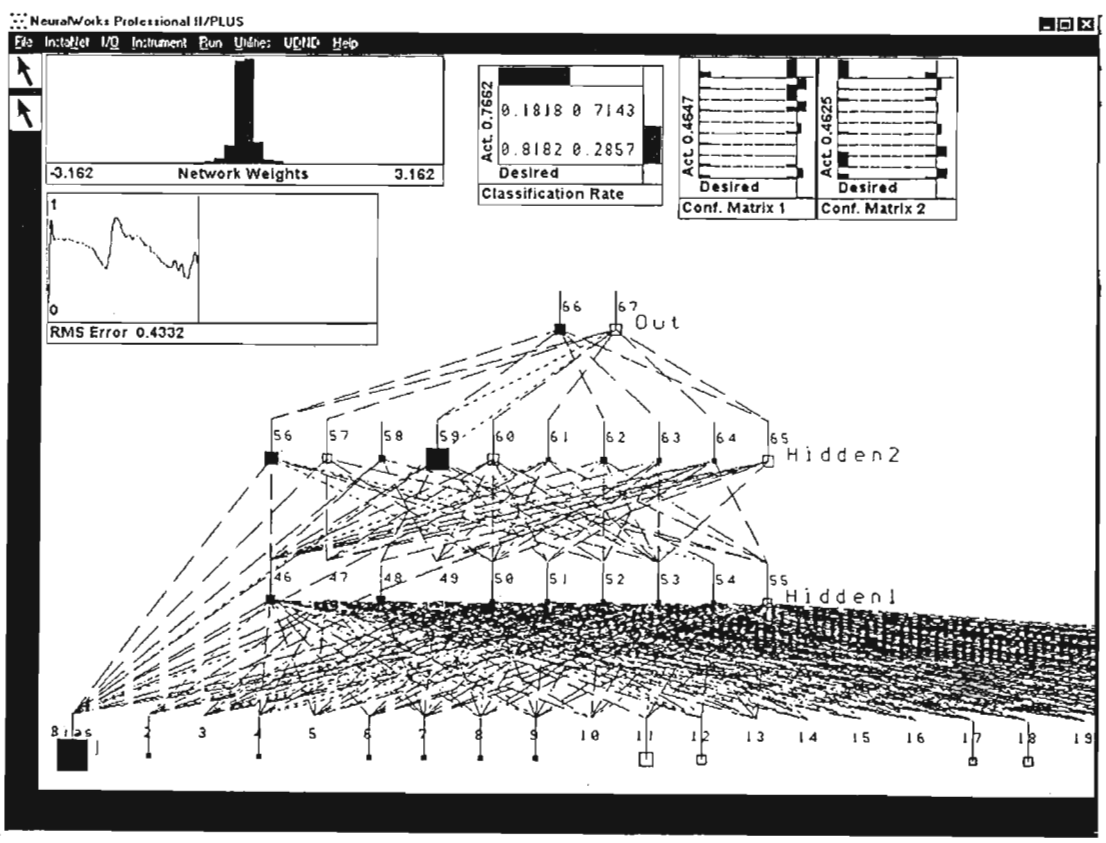


Figure 6.14 *NeuralWorks Professional II+* graphical user interface

There are several instruments available in *NeuralWorks Professional II+* which help in training and testing neural networks.

RMS error

- used in prediction/regression analysis
- indicates the error root-mean-square error of all the neurons in the output layer
- gives a measure of the closeness of fit of desired to actual neural network response

Weight histogram

- used to check condition and spread of neural weights
- provides a measure of the entire network
- the height of the bars represents the number of weights in each weight category
- for a well-trained, non-saturated neural network, the weights should be ideally Gaussian distributed about zero or evenly distributed across the weight range

Confusion matrix

- used in classification/ranking problems
- indicates the correlation of actual to desired to actual results for each category
- the confusion matrix is a analogue of a scatter-plot where the height of the bars in each "bin" represent the number of "hits" within that bin
- ideally the bars should line up along the diagonal
- the numerical figure indicates the correlation for each matrix

Classification rate matrix

- used in classification/ranking problems
- similar to the confusion matrix, it provides a measure of correlation for the entire network
- the figures on the diagonal represent the fraction of correctly classified training or test cases
- the figure off-diagonal represent the fraction of misclassified classes
- it can provide a measure of which classifications are overlapping or difficult to separate

6.6 Practical issues in developing back-propagation neural networks

Number of input neurons

The type of data used determines the number of input neurons. In other words, the more feature descriptors that are used, the larger the input layer will be. Generally it is better to commence with more feature descriptors and selectively delete those which have little or no effect on learning.

Number of hidden neurons

The choice of this number is one of the most debated topics in neural networks. Fundamentally the number of hidden neurons depends strongly on the type of data used. That is, if the data is simple (e.g. easily separable) then fewer neurons are required than in the case of complex data sets. Basically, each hidden neuron provides an extra hyperplane to separate regions in multi-dimensional hyperspace. If the contour of distinction between categories, for example, is complex more hyperplanes will be needed to define it.

Widrow and Stearns, (1985) recommend that the number of hidden neurons be related to the number of training samples available. They states that the ratio of the number of weights to the number of training samples should be around 10% for standard back-propagation neural networks. In most instances, the type of problem defines the number of input and output neurons, so calculation of the number of weights in a back-propagation neural network is simple.

Several heuristics exist, but do not take into account a measure of data complexity. A better method is to use the designed experiment technique. This allows a scientific, rigorous approach to choosing this number. A set of experiments is designed, neural networks constructed and tested with different numbers of hidden neurons and the optimal configuration selected. This method was used in neural network development in this thesis.

Number of hidden layers

In general it is advisable to start with a single hidden layer in the back-propagation neural network and only progress to two hidden layers when training became difficult. The need for three hidden layers is rarely found.

Choice of learning paradigm

With back-propagation, it is usually best to use the extended-delta-bar-delta heuristic for learning. It alleviates the need to set the momentum and learning coefficient terms interactively, by automatically selecting these parameters during training. The extended delta-bar-delta heuristic was used in all back-propagation neural networks designed in this thesis.

Epoch size

The number of training data passes that are allowed before cumulative update of weights is termed the epoch size. Ideally this should be as large as the number of training samples, however owing to the extensive training set available, the epoch size was limited, usually between 250 and 1000. The limit was required to improve learning speed and allow the influence of training samples with a lower rate of occurrence to influence the weights. If the full epoch size had been chosen, the rarer data samples would have been eliminated in the generalisation that occurs.
Optimal epoch size can be chosen using a designed experiment method in which various sizes are tested for the highest correlation value. This is a more thorough approach yet had little effect on the large data set available for training.

Weight initialisation

Weights were initialised randomly within a small range around zero. Typically the values were between -0.2 and +0.2.

6.7 Preliminary Neural Network Combinations

A number of initial neural networks were trained in order to see which compound groupings gave the best results. That is, whether or not individual recognition of each compound was possible at the outset, and whether grouping the compounds according to their classes was necessary to improve classification. If the latter was true, it had to be established which class groupings gave the best results. The following results demonstrate this.

6.7.1 Classification of Whole Data set as Individuals

Neural network: Class01 In this preliminary case, recognition of each individual compound was attempted. This network was trained to identify each of the 69 pure compounds using their 10 ppm binned ^{13}C NMR spectral data as inputs. The target output for the neural network was a binary one-of-*n* encoded representation of each individual compound, i.e.

compound 1 = 0 0 .....0 0 1
compound 2 = 0 0 .....0 1 0
...
compound 69 = 1 0 .....0 0 0

The binary one-of-*n* encoding method ensures maximal differentiation between each classification group since each target vector is orthogonal with maximal Hamming distance in 69-dimensional hyperspace.

After training, the neural network identified 57% of the compounds correctly when tested on the training dataset of pure compounds. When tested on the impure spectra of the pure compounds used in training, the number of correctly identified compounds decreased to 44%. This poor result indicated that the blind application of neural networks most often yields inconclusive results. A more thorough approach to neural network design was required.

6.7.2 Classification of Data set into various groupings

To classify the compounds according to their groupings, the training set consisted of the 10 ppm binned ^{13}C NMR spectral data of the 69 pure compounds and the test set consisted of the 10 ppm binned ^{13}C NMR spectral data of the 66 impure compounds. As the protolimonoids, glabretal-type protolimonoids, dammaranes, masticadienoic acids and sterols are all different types of triterpenoids, various combinations were possible and tried in order to achieve the overall best classification results.

Neural Network: Class02: the dataset was grouped into the limonoids, protolimonoids, glabretal-type protolimonoids, dammaranes, triterpenoids (masticadienoic acids, sterols) and flavonoid/coumarins. Thus the neural network was trained to place all the compounds into one of these classes.

Neural Network: Class03: the dataset was grouped into the limonoids, protolimonoids, dammaranes, triterpenoids (glabretal-type protolimonoids, masticadienoic acids, sterols) and flavonoid/coumarins.

Neural Network: Class04: the dataset was grouped into the limonoids, triterpenoids (protolimonoids, glabretal-type protolimonoids, masticadienoic acids, sterols), dammaranes and flavonoid/coumarins.

Neural Network: Class06: the dataset was grouped into the limonoids, protolimonoids (including the glabretal-type protolimonoids), triterpenoids (masticadienoic acids, sterols and dammaranes) and flavonoid/coumarins.

Neural Network: Class07: the dataset was grouped into the limonoids, triterpenoids (protolimonoids, glabretal-type protolimonoids, masticadienoic acids, sterols and dammaranes) and flavonoid/coumarins.

The results of the various network combinations are summarised in the table below:

Table 6.1

Neural Network	Classes						Results	
	limonoids	protolimo-noids	glabretal-type protolimo noids	damm-aranes	triter-penoids	coumarin /flavo noids	Training set	Test set
Neural Network: Class02	✓	✓	✓	✓	✓	✓	97%	69%
Neural Network: Class03	✓	✓		✓	✓	✓	92%	85%
Neural Network: Class04	✓			✓	✓	✓	98%	84%
Neural Network: Class06	✓	✓			✓	✓	98%	77%
Neural Network: Class07	✓				✓	✓	100%	85%

It was evident from the results of the above table that the best classification results were obtained when only the three categories: limonoids, triterpenoids (which included the protolimonoids, glabretal-type protolimonoids, dammaranes, masticadienoic acids and sterols) and flavonoid/coumarins were used.

In order to assess the merits of clumping dissimilar data into an "other" category, the following groupings were attempted:

Neural Network: Class08: the dataset was grouped into the limonoids or "other"(which included the protolimonoids, glabretal-type protolimonoids, masticadienoic acids, sterols, dammaranes and flavonoid/coumarins.)

Neural Network: Class09: the dataset was grouped into the limonoids, protolimonoids (which included the glabretal-type protolimonoids) or "other"(which included the masticadienoic acids, sterols, dammaranes and flavonoid/coumarins.)

The results are summarised in the table below:

Table 6.2

Neural Network	Classes			Results	
	limonoids	protolimonoids	other	Training set	Test set
Neural Network: Class08	✓		✓	98%	89%
Neural Network: Class09	✓	✓	✓	92%	85%

The results show the ability of *Predict* to correctly associate and cluster quite dissimilar compounds into three fundamental supersets.

Lastly, in order to assess whether the binning size chosen was correct, two networks were trained to classify the whole data set into limonoids, protolimonoids, glabretal-type protolimonoids, triterpenoids (which included the sterols, dammaranes and masticadienoic acids) and "other"(flavonoids/coumarins) using different binning sizes:

Neural Network: Class49: 5 ppm binning was used
Neural Network: Class48: 10 ppm binning was used

The results are summarised in the table below:

Table 6.3

Neural Network	Classes					Results	
	limo-noids	protolimo-noids	glabretal type protolim-onoids	triter-penoids	coumarin/ flavonoids	Training set	Test set
Neural Network: class49 (5ppm)	✓	✓	✓	✓	✓	95%	78%
Neural Network: class48 (10ppm)	✓	✓	✓	✓	✓	93%	75%

There was clearly an improved performance as a result of using 5 ppm binning, thus implying that the 5 ppm bin width should be the width of choice.

6.8 Hybrid neural network architecture

Following the results of the preliminary neural networks, it was decided that a hybrid neural network structure would be more efficient in the differentiation between limonoids (including limonoids as individuals and types), protolimonoids, glabretal-type protolimonoids and "other" compounds (flavonoids/coumarins). In the hybrid case, the problem would be partitioned so that two decision levels in a tree-type structure would provide coarse and fine classification respectively. Each decision level would use a number of neural networks and would progress to a tighter classification at each level down the tree. The benefit of this approach is that a wide range of products from many classes could be accommodated within a single decision strategy, and could be expanded with ease. It also improved the overall robustness of the classification decision by using the combined and progressive decision of a number of neural networks in a "panel-of-experts" configuration.

The first neural network (NN1) would discriminate between limonoids, triterpenoids (which included the protolimonoids, glabretal-type protolimonoids, plant sterols, dammaranes and masticadienoic acids) and "other" (flavonoids/coumarins). The separated triterpenoids would then undergo either one or the other of two parallel neural networks, namely NN2(a) and NN2(b) depending on the results of NN1. NN2(a) would recognise the protolimonoids (which included the glabretal type protolimonoids) from the rest of the triterpenoids, while NN2(b) would recognise three groups of compounds, namely: the protolimonoids, glabretal type protolimonoids and the rest of the triterpenoids. Lastly, the limonoids would be treated by one or the other of two neural networks, namely NN3(a) and NN3(b). NN3(a) would be trained to recognise each individual limonoid in the dataset while NN3(b) would be able to group the limonoid dataset into three main groupings i.e. no oxidation (no carbonyl groups), α,β-unsaturation in ring A, and keto groups at C-9, C-15 or C-3 .

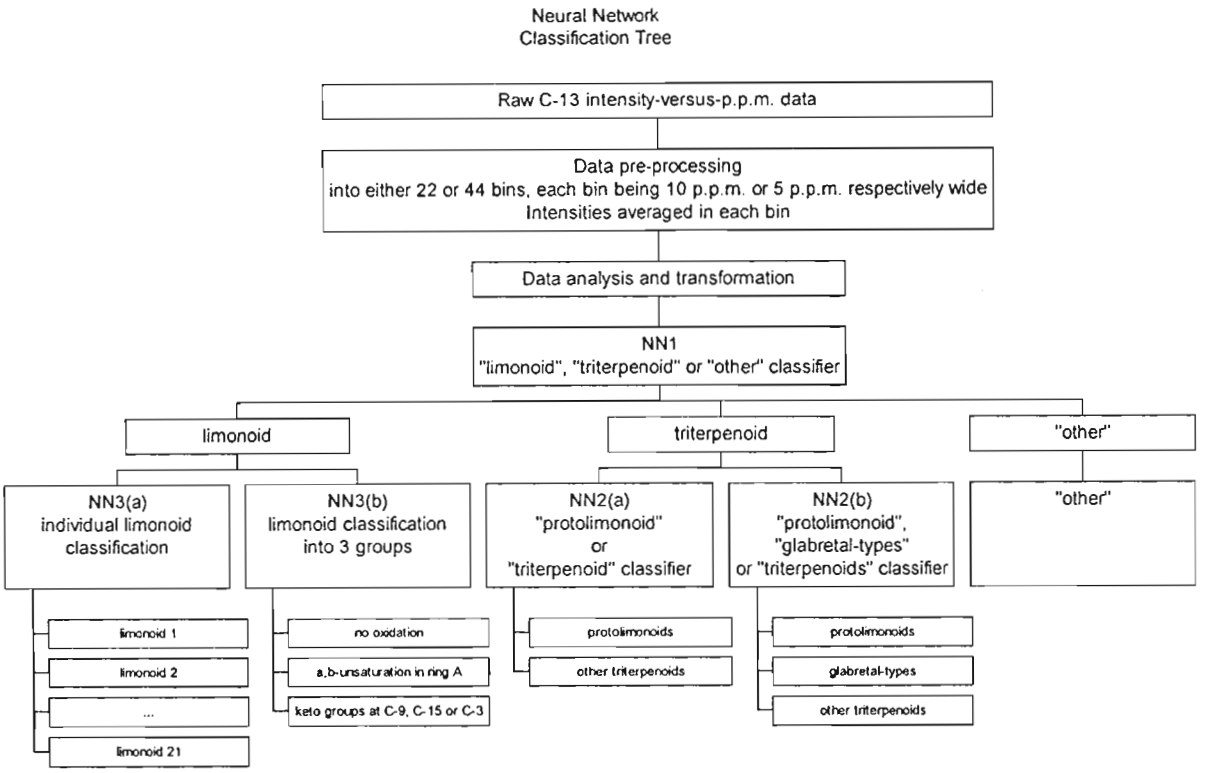


Figure 6.15 Hybrid Neural Network Structure

6.9 Classification of Whole Data set into limonoids, triterpenoids, or "other" (NN1)

Unlike Munk *et al.*, (1996) and Lohninger *et al.*, (1992) who used the terpene skeleton as a substructural feature, limonoids, protolimonoids, plant sterols, dammaranes and masticadienoic acids are all triterpenoid based structures so that discrimination on this criterion was impossible. Substructural features such as alcohols, alkenes, esters, carboxylic acids, saturated and unsaturated ketones, carbomethoxy groups, formate groups, lactones, ethers and acetate groups are frequently encountered in all of them as well as in the flavonoid/coumarin class which are structurally different compounds. This further complicates classification. The presence of a number of substructural features within each 30-40 carbon compound, together with extraneous peaks due to impurities as well as

stereoisomeric mixtures and mixtures of esters increases the complexity of the problem. Thus, the initial network (NN1) was designed to discriminate between "limonoids" (triterpenoid-type structure + furan ring), "triterpenoids" and the frequently encountered flavonoid/coumarin ("other") class which was added to show that the network could tolerate totally different spectral profiles outside of the main focus group. The binning technique is a robust method of generalising the overall spectral pattern while at the same time allowing it to retain characteristic features. Limonoids are easily distinguishable from the other two categories by their characteristic furan ring carbon signals at approximately δ142, δ140, δ125, δ110. Thus 10 ppm binning should show enhanced signals in the 3 bins marked δ140-150, δ120-130 and δ110-120, while the 5 ppm binning would show enhanced signals in the 3 bins marked δ140-145, δ125-130 and δ110-115. The neural network was trained to associate each input pattern from both the 5 ppm and 10 ppm binned ^{13}C NMR spectra to its corresponding compound.

6.9.1 Classification of Whole Data set into limonoids, triterpenoids, or "other" using 5ppm binning

Three different neural networks were attempted using 5 ppm binning, namely **Neural Network: Class37**, **Neural Network: 38** and **Neural Network: 39**.

Neural Network: Class37: Since the 135 compounds of the whole dataset comprised: **57** limonoids (28 pure and 29 impure), **61** triterpenoids (30 pure and 31 impure) and only **17** "others" which was the flavonoid/coumarin set (11 pure and six impure), it was necessary to increase the number of the "other" category compounds so as to increase its significance and representation. This was important since the neural network's generalising ability is based on the frequency of occurrence of each category. Poor representation of any category reduces its significance in learning. This was achieved by duplicating some of the members of the flavonoid/coumarin dataset randomly , thus making its representation **46,** which was more comparable to the **57** limonoids and **61** triterpenoids. The whole dataset then consisted of a total of **164** records. The neural network was trained on 132 of both the impure and pure compounds of the whole dataset, while 32 compounds

(16 pure and 16 impure) were taken for the unseen test set. Representations of each class were proportionately distributed in both the training and the test sets.

Neural Network: Class38: In this network, only the 69 pure compounds were used for both training and testing. Thus 53 pure compounds were used as the training set and 16 pure compounds were used as the test set. There was no duplication of the flavonoid/coumarin class and representations of each class were proportionately distributed in both the training and the test sets.

Neural Network: Class39: In this network, the 69 pure compounds were used as the training set, and the 66 impure compounds were used as the test set. Again, there was no duplication of the flavonoid/coumarin class and representations of each class were proportionately distributed in both the training and test sets.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.4

	Neural Network	NN1		NN1		NN1	
General Parameters	Name	Neural Network: Class37		Neural Network: Class38		Neural Network: Class39	
	Training set (seen)	132 pure & impure		53 pure		69 pure	
	Test set (unseen)	16 pure & 16 impure		16 pure		66 impure	
	Bin size	5 ppm		5 ppm		5 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	6		4		5	
	Hidden Neurons	0		2		2	
	Output Neurons	3		3		3	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	44	1	44	1	44	1
	Active fields	6	1	4	1	5	1
	Total Transforms	158	3	154	3	157	3
	Active Transforms	6	3	4	3	5	3
	Input data transformations	fuzzy, linear, log		fuzzy, linear, inv. power 4,		fuzzy, linear, log	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.9.2 Classification of Whole Data set into limonoids, triterpenoids, or "other" using Learning Vector Quantisation (LVQ)

By means of comparison, LVQ was also applied to the 5 ppm binned data in **Neural Network: Class37, Neural Network: 38** and **Neural Network: Class39**. The network structures are shown in the table below:

Table 6.5

	Neural Network: Class37	Neural Network: Class38	Neural Network: Class39
Input neurons	44	44	44
Kohonen neurons	13	6	6
Output neurons	3	3	3

6.9.3 Classification of Whole Data set into limonoids, triterpenoids, or "other" using Back Propagation Neural Networks

Also, as a means of comparison, back propagation neural networks were applied to the 5 ppm binned data in **Neural Network: Class37**, **Neural Network: Class38** and **Neural Network: Class39**. The network structures are shown in the table below:

Table 6.6

	Neural Network: Class37	Neural Network: Class38	Neural Network: Class39
Input neurons	44	44	44
Hidden neurons	20	20	30
Output neurons	3	3	3

6.9.4 Classification of Whole Data set into limonoids, triterpenoids, or "other" using 10ppm binning

Three different neural networks were attempted using 10 ppm binning, namely **Neural Network: Class60**, **Neural Network: Class29** and **Neural Network: Class59**.

Neural Network: Class60: The same dataset as in **Neural Network: Class37** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class29: The same dataset as in **Neural Network: Class38** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class59: The same dataset as in **Neural Network: Class39** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.7

	Neural Network	NN1		NN1		NN1	
General Parameters	Name	Neural Network: Class60		Neural Network: Class29		Neural Network: Class59	
	Training set (seen)	132 pure & impure		53 pure		69 pure	
	Test set (unseen)	16 pure & 16 impure		16 pure		66 impure	
	Bin size	10 ppm		10 ppm		10 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	11		7		8	
	Hidden Neurons	0		0		0	
	Output Neurons	3		3		3	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	22	1	22	1	22	1
	Active fields	10	1	7	1	8	1
	Total Transforms	81	3	80	3	80	3
	Active Transforms	11	3	7	3	8	3
	Input data transformations	fuzzy, inverse, inverse 4th root, linear, log		fuzzy, inverse, linear		fuzzy, inverse, linear	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.10 Classification of triterpenoids dataset into protolimonoids and triterpenoids (NN2(a))

The triterpenoids which were separated from the limonoids and "other" (flavonoids/coumarins) in NN1 could then be dealt with in this network or NN2(b). This network was trained to discriminate between the protolimonoids (which included the glabretal type protolimonoids) and the rest of the triterpenoids (which included the dammaranes, masticadienoic acids and sterols) in the triterpenoid dataset. This was difficult as the protolimonoids fall into the general category of triterpenoids.

6.10.1 Classification of triterpenoids dataset into protolimonoids and triterpenoids using 5 ppm binning

Three different neural networks were attempted using 5 ppm binning, namely **Neural Network: Class24**, **Neural Network: Class30** and **Neural Network: Class16**.

Neural Network: Class24: The dataset consisted of **58** triterpenoids (30 pure and 28 impure). The neural network was trained on **46** impure and pure compounds of the dataset, while 12 compounds (6 pure and 6 impure) were taken for the unseen test set. Representations of each class (namely, triterpenoid or protolimonoid) were proportionately distributed in both the training and the test sets.

Neural Network: Class30: In this network, only the 30 pure triterpenoids were used for both training and testing. Thus 24 pure compounds were used as the training set and 6 pure compounds were used as the test set. Again, representations of each class were proportionately distributed in both the training and the test sets.

Neural Network: Class16: In this network, the 30 pure compounds were used as the training set, and the 28 impure compounds were used as the test set. Again, representations of each class were proportionately distributed in both the training and test sets.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.8

	Neural Network	NN2(a)		NN2(a)		NN2(a)	
General Parameters	Name	**Neural Network: Class24**		**Neural Network: Class30**		**Neural Network: Class16**	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure and impure		6 pure		28 impure	
	Bin size	5 ppm		5 ppm		5 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	7		5		7	
	Hidden Neurons	0		0		0	
	Output Neurons	2		2		2	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	44	1	44	1	44	1
	Active fields	7	1	5	1	6	1
	Total Transforms	125	2	118	2	124	2
	Active Transforms	7	2	5	2	7	2
	Input data transformations	fuzzy, linear		fuzzy, linear		fuzzy, inverse, linear, inv. 4^{th} root	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.10.2 Classification of triterpenoids dataset into protolimonoids and triterpenoids using Learning Vector Quantisation (LVQ)

By means of comparison, LVQ was also applied to the 5 ppm binned data in **Neural Network: Class24**, **Neural Network: Class30** and **Neural Network: Class16**. The network structures are shown in the table below:

Table 6.9

	Neural Network: Class24	Neural Network: Class30	Neural Network: Class16
Input neurons	44	44	44
Kohonen neurons	4	3	4
Output neurons	2	2	2

6.10.3 Classification of triterpenoids dataset into protolimonoids and triterpenoids using Back Propagation Neural Networks

Also, as a means of comparison, back propagation neural networks were applied to the 5 ppm binned data in **Neural Network: Class24**, **Neural Network: Class30** and **Neural Network:Class16**. The network structures are shown in the table below:

Table 6.10

	Neural Network: Class24	Neural Network: Class30	Neural Network: Class16
Input neurons	44	44	44
Hidden neurons	20	20	20
Output neurons	2	2	2

6.10.4 Classification of triterpenoids dataset into protolimonoids and triterpenoids using 10ppm binning

Three different neural networks were attempted using 10 ppm binning, namely **Neural Network: Class 20, Neural Network: Class33 and Neural Network: Class12.**

Neural Network: Class20: The same dataset as in **Neural Network: Class24** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class33: The same dataset as in **Neural Network: Class30** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class12: The same dataset as in **Neural Network: Class16** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.11

	Neural Network	NN2(a)		NN2(a)		NN2(a)	
General Parameters	Name	**Neural Network: Class20**		**Neural Network: Class33**		**Neural Network: Class12**	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure & impure		6 pure		28 impure	
	Bin size	10 ppm		10 ppm		10 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	15		7		8	
	Hidden Neurons	0		0		0	
	Output Neurons	2		2		2	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	22	1	22	1	22	1
	Active fields	12	1	7	1	7	1
	Total Transforms	73	2	70	2	70	2
	Active Transforms	15	2	7	2	8	2
	Input data transformations	fuzzy, linear, log, tanh, inv. power 2		fuzzy, linear, inv., inv. 4^{th} root, log		fuzzy, inverse, linear, log	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.10.5 Classification of triterpenoids dataset into protolimonoids and triterpenoids using 15 ppm binning

Three different neural networks were attempted using 15 ppm binning, namely **Neural Network: Class 25, Neural Network: Class35** and **Neural Network: Class17**.

Neural Network: Class25: The same dataset as in **Neural Network: Class20** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

Neural Network: Class35: The same dataset as in **Neural Network: Class33** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

Neural Network: Class17: The same dataset as in **Neural Network: Class12** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.12

	Neural Network	NN2(a)		NN2(a)		NN2(a)	
General Parameters	Name	Neural Network: Class25		Neural Network: Class35		Neural Network: Class17	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure & impure		6 pure		28 impure	
	Bin size	15 ppm		15 ppm		15 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	8		6		5	
	Hidden Neurons	1		2		0	
	Output Neurons	2		2		2	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	15	1	15	1	15	1
	Active fields	5	1	5	1	5	1
	Total Transforms	54	2	50	2	50	2
	Active Transforms	8	2	6	2	5	2
	Input data transformations	fuzzy, linear, inv.4th root, log, lnx/(1-x)		fuzzy, linear, inv.4th root		fuzzy, 4th root, inverse	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.11 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids (NN2(b))

The triterpenoids which were separated from the limonoids and "other" (flavonoids/coumarins) in NN1 could then be dealt with in this network or in NN2(a) as already shown. This network was trained to discriminate between the protolimonoids, the glabretal-type protolimonoids and the rest of the triterpenoids

(which included the dammaranes, masticadienoic acids and sterols) in the triterpenoid dataset. This is difficult, as there is no clearly defined distinction between the three classes since the protolimonoids and glabretal-type protolimonoids fall into the general category of triterpenoids. Also, as previously mentioned at the beginning of this chapter, two dysoxylic acids were classified as glabretal type protolimonoids although they are really borderline between the glabretal-type protolimonoid class and the triterpenoid class as they have the glabretal nucleus but the masticadienoic acid side chain.

6.11.1 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using 5 ppm binning

Three different neural networks were attempted using 5 ppm binning, namely **Neural Network: Class27**, **Neural Network: 31** and **Neural Network: 52**.

Neural Network: Class27: The dataset consisted of **58** triterpenoids (30 pure and 28 impure). The neural network was trained on **46** impure and pure compounds of the dataset, while 12 compounds (6 pure and 6 impure) were taken for the unseen test set. Representations of each class (namely, triterpenoid, protolimonoid or glabretal-type protolimonoid) were proportionately distributed in both the training and the test sets.

Neural Network: Class31: In this network, only the 30 pure triterpenoids were used for both training and testing. Thus 24 pure compounds were used as the training set and 6 pure compounds were used as the test set. Again, representations of each class were proportionately distributed in both the training and the test sets.

Neural Network: Class52: In this network, the 30 pure compounds were used as the training set, and the 28 impure compounds were used as the test set. Again, representations of each class were proportionately distributed in both the training and test sets.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.13

	Neural Network	NN2(b)		NN2(b)		NN2(b)	
General Parameters	Name	Neural Network: Class27		Neural Network: Class31		Neural Network: Class52	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure & impure		6 pure		28 impure	
	Bin size	5 ppm		5 ppm		5 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	8		6		8	
	Hidden Neurons	0		0		0	
	Output Neurons	3		3		3	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	44	1	44	1	44	1
	Active fields	8	1	6	1	7	1
	Total Transforms	126	3	118	3	124	3
	Active Transforms	8	3	6	3	8	3
	Input data transformations	linear, inv.4th root, inv.2nd root, inv. power 4, inv. power 2		inverse, linear, tanh		fuzzy, inverse, linear, log, tanh,	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.11.2 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using Learning Vector Quantisation (LVQ)

By means of comparison, LVQ was also applied to the 5 ppm binned data in **Neural Network: Class27**, **Neural Network: 31** and **Neural Network: 52**. The network structures are shown in the table below:

Table 6.14

	Neural Network: Class27	Neural Network: Class31	Neural Network: Class52
Input neurons	44	44	44
Kohonen neurons	6	2	3
Output neurons	3	3	3

6.11.3 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using Back Propagation Neural Networks

Also, as a means of comparison, back propagation neural networks were applied to the 5 ppm binned data in **Neural Network: Class27**, **Neural Network: 31** and **Neural Network: 52**. The network structures are shown in the table below:

Table 6.15

	Neural Network: Class27	Neural Network: Class31	Neural Network: Class52
Input neurons	44	44	44
Hidden neurons	20	20	20
Output neurons	3	3	3

6.11.4 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using 10 ppm binning

Three different neural networks were attempted using 10 ppm binning, namely **Neural Network: Class 26**, **Neural Network: 32** and **Neural Network: 54**.

Neural Network: Class26: The same dataset as in **Neural Network: Class27** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class32: The same dataset as in **Neural Network: Class31** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

Neural Network: Class54: The same dataset as in **Neural Network: Class52** above was used for both training and testing, the only difference being the 10 ppm binning as opposed to the 5 ppm binning.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.16

	Neural Network	NN2(b)		NN2(b)		NN2(b)	
General Parameters	Name	Neural Network: Class26		Neural Network: Class32		Neural Network: Class54	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure & impure		6 pure		28 impure	
	Bin size	10 ppm		10 ppm		10 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Mod. Noisy data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Exhaustive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	9		7		7	
	Hidden Neurons	0		3		0	
	Output Neurons	3		3		3	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	22	1	22	1	22	1
	Active fields	9	1	7	1	7	1
	Total Transforms	73	3	70	3	70	3
	Active Transforms	9	3	7	3	7	3
	Input data transformations	fuzzy, inverse, linear, inv.4th root		fuzzy, inverse, linear, inv. power 2		fuzzy, inverse, linear, 4th root, log	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.11.5 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using 15 ppm binning

Three different neural networks were attempted using 15 ppm binning, namely **Neural Network: Class 28**, **Neural Network: 34** and **Neural Network: 53**.

Neural Network: Class28: The same dataset as in **Neural Network: Class26** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

Neural Network: Class34: The same dataset as in **Neural Network: Class32** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

Neural Network: Class53: The same dataset as in **Neural Network: Class54** above was used for both training and testing, the only difference being the 15 ppm binning as opposed to the 10 ppm binning.

The parameters that were used in *Predict* are detailed in the tables below:

Table 6.17

	Neural Network	NN2(b)		NN2(b)		NN2(b)	
General Parameters	Name	Neural Network: Class28		Neural Network: Class34		Neural Network: Class53	
	Training set (seen)	46 pure & impure		24 pure		30 pure	
	Test set (unseen)	12 pure & impure		6 pure		28 impure	
	Bin size	15 ppm		15 ppm		15 ppm	
	Problem Type	Classification		Classification		Classification	
	Noise level	Clean data		Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive		Comprehensive	
	Network Search	Comprehensive		Comprehensive		Comprehensive	
Network Structure	Network Type	feedforward		feedforward		feedforward	
	Input Neurons	10		6		8	
	Hidden Neurons	0		0		0	
	Output Neurons	3		3		3	
	Direct Connections	yes		yes		yes	
	Cascade Connections	yes		yes		yes	
Data Transforms		In	Out	In	Out	In	Out
	Total fields	15	1	15	1	15	1
	Active fields	8	1	6	1	8	1
	Total Transforms	54	3	50	3	50	3
	Active Transforms	10	3	6	3	8	3
	Input data transformations	fuzzy, inverse, linear, 4^{th} root, log		fuzzy, linear, inv.4^{th} root, log, 2^{nd} root		fuzzy, inverse, linear, inv.4^{th} root, log, loglog	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh		tanh	
	Output Layer	softmax		softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate		ave. class. rate	

6.12 Classification of limonoids as individuals (NN3(a))

This neural network dealt with the limonoids that were separated from the triterpenoids and "other" classes in the original NN1. The neural network was trained to recognise each of 21 individual pure pre-identified limonoids within the limonoid data set. The network was then tested on 22 impure limonoids. Again, the

whole spectrum is important as the limonoids can vary considerably in all regions since they may have rings A and/or B and /or D opened, extra side-chains, extra acetate groups, saturated and unsaturated ketones, ethers, lactones and formates. Some of the ekebergin type limonoids were mixtures of esters that meant two peaks for a single carbon atom within the molecule. This eliminated the possibility of peak picking and reinforced the binning technique.

6.12.1 Classification of limonoids as individuals using 5 ppm and 10 ppm binning

Neural Network: Class40: This network was trained on the 5 ppm binned data of 21 limonoids in the pure limonoid dataset. The network was trained to recognise each of the 21 limonoids as different individuals and thus the 22 limonoids in the test set had to be impure limonoids of the 21 in the training set.

Neural Network: Class13: This network was trained and tested on the same data as Class40 above, the only difference being that the data was 10 ppm binned as opposed to 5 ppm binned.

The parameters that were used in *Predict* are shown in the table below:

Table 6.18

	Neural Network	NN3		NN3	
General Parameters	Name	Neural Network: Class40		Neural Network: Class13	
	Training set (seen)	pure 21		21 pure	
	Test set (unseen)	impure 22		22 impure	
	Bin size	5 ppm		10 ppm	
	Problem Type	Classification		Classification	
	Noise level	Clean data		Clean data	
	Data Transforms	Comprehensive		Comprehensive	
	Variable Selection	Comprehensive		Comprehensive	
	Network Search	Comprehensive		Exhaustive	
Network Structure	Network Type	feedforward		feedforward	
	Input Neurons	6		7	
	Hidden Neurons	0		0	
	Output Neurons	22		22	
	Direct Connections	yes		yes	
	Cascade Connections	yes		yes	
Data Transforms		In	Out	In	Out
	Total fields	44	1	22	1
	Active fields	6	1	7	1
	Total Transforms	129	22	69	22
	Active Transforms	6	22	7	22
	Input data transformations	fuzzy, inverse, linear		fuzzy, inverse, linear, 4^{th} root, tanh,	
	Output data transformations	binary 1-of-*n*		binary 1-of-*n*	
Variable Selection	Model	multiple regression		multiple regression	
	Evaluation Technique	genetic algorithm		genetic algorithm	
Learning Rule	Type	adaptive gradient		adaptive gradient	
Transfer Function	Hidden Layer	tanh		tanh	
	Output Layer	softmax		softmax	
Evaluation Function	Type	ave. class. rate		ave. class. rate	

6.12.2 Classification of limonoids as individuals using Learning Vector Quantisation (LVQ)

By means of comparison, and owing to the poor results achieved in *Predict*, LVQ was also applied to the 5 ppm and 10 ppm binned data in **Neural Network: Class40** and **Neural Network: Class13** respectively. The network structures are shown in the table below:

Table 6.19

	Neural Network: Class40	Neural Network: Class13
Input neurons	44	22
Kohonen neurons	63	63
Output neurons	21	21

6.12.3 Classification of limonoids as individuals using Back Propagation Neural Networks

Also, as a means of comparison, back propagation neural networks were applied to the 5 ppm and 10 ppm binned data in **Neural Network: Class40** and **Neural Network: Class13** respectively. The network structures are shown in the table below:

Table 6.20

	Neural Network: Class40	Neural Network: Class13
Input neurons	44	22
Hidden neurons	20	20
Output neurons	21	21

6.13 Classification of limonoids into 3 Groups using 5ppm binning (NN3(b))

This neural network, **Neural Network: Class 56** dealt with the limonoids which were separated from the triterpenoids and "other" classes in the original NN1. The neural network was trained to categorise each of the 21 individual pure pre-identified limonoids within the limonoid data set into three groups or classes, according to three criteria:

- no oxidation i.e.: no carbonyl groups
- α,β-unsaturation in ring A
- keto groups at C-9, C-15 or C-3.

The dataset consisted of 21 pure limonoids, 22 impure limonoids and 13 extra unseen limonoids that were a mixture of both pure and impure unseen limonoids. The network was trained on the 5 ppm binned data of the 21 pure limonoids and tested on the 5 ppm binned data of the 22 impure limonoids which were one of the 21 pure limonoids in the training set. The 5 ppm binned data of the 13 unseen limonoids were used as an unseen test set.

The parameters that were used in *Predict* are shown in the table below:

Table 6.21

	Neural Network	NN3(b)	
	Name	Neural Network: Class56	
	Training set (seen)	pure 22	
	Test set (unseen)	impure 22	
	Test set (unseen)	13 pure & impure	
	Bin size	5 ppm	
General Parameters	Problem Type	Classification	
	Noise level	Mod. noisy data	
	Data Transforms	Comprehensive	
	Variable Selection	Comprehensive	
	Network Search	Comprehensive	
Network Structure	Network Type	feedforward	
	Input Neurons	3	
	Hidden Neurons	0	
	Output Neurons	3	
	Direct Connections	yes	
	Cascade Connections	yes	
Data Transforms		In	Out
	Total fields	44	1
	Active fields	3	1
	Total Transforms	129	3
	Active Transforms	3	3
	Input data transformations	fuzzy, inverse, logical	
	Output data transformations	binary 1-of-*n*	
Variable Selection	Model	multiple regression	
	Evaluation Technique	genetic algorithm	
Learning Rule	Type	adaptive gradient	
Transfer Function	Hidden Layer	tanh	
	Output Layer	softmax	
Evaluation Function	Type	ave. class. rate	

6.14 Sensitivity Analysis

Sensitivity analysis allows one to determine the effect that a small change in an input value will have on the output value(s), and to rank the input fields according to this sensitivity. In mathematical terms, the output of Sensitivity analysis is a matrix of partial derivatives of output variables with respect to input variables.

Sensitivity analysis can provide good insights into the model. High sensitivity to a particular input that is likely to fluctuate around its current value may cause the output to be questionable. If the field values could be controlled, sensitivity analysis reveals which fields must be changed and in which direction to change them, so as to achieve the desired model output.

Highly sensitive fields may not necessarily be important fields. One way to visualise this is to think of a gently sloping corrugated roof. This can be described by a function which maps an (x,y) co-ordinate to a height on the roof. The function's x-co-ordinate is measured in a direction across the ridges, and its y-co-ordinate is measured in a direction along the ridges. In general, the height is very sensitive to small changes in x but not very sensitive to small changes in y. However, y is the more important variable. An approximate value for the height of the roof can be found using just the y-co-ordinate, but the x-co-ordinate by itself is fairly randomly related to height. Sensitivity analysis outputs a matrix of values for each record. Each entry in that matrix represents the sensitivity of a particular output field with respect to a particular input field.

6.15 Contribution Analysis

Contribution analysis rates the input fields for a given input record based on the contribution they have made to the outcome. This contribution can be thought of as the additional information a field has brought to the prediction as compared to if it was missing. Contribution analysis assigns each field a value between -100 and 100. A value of 0 is assigned to missing fields and inactive fields. A value of 100 is assigned to the field that had the most pronounced effect on the output for the current record. A value of -100 is assigned to the field that has the least pronounced effect on the output for the current record.

6.16 Summary

This chapter began by detailing the source of all the natural products used in the data set, following which the binning technique, which was the pre-processing method used to make the raw ^{13}C NMR data more acceptable, was discussed in detail. The parameters used in *Predict* which were tabulated in the experimental tables for each neural network, were discussed in detail and which included the input and output transformation variables, the structure of the neural network, the transfer function of the hidden and output layers of the neural network, the learning rule and lastly the evaluation technique.

A number of preliminary natural products groupings were outlined, on which neural networks were run in order to see which groupings gave the best results. Finally the hybrid neural network architecture was proposed and *Predict* run on all the data and their various bin widths. Since the best neural network results were achieved using the 5 ppm bin width, LVQ and back-propagation neural networks were also run on this 5 ppm binned data. Finally, sensitivity analysis was discussed as a technique to evaluate the stability of the neural network structures. Contribution analysis, which gave an insight into the important data input field contributors, was also considered.

Chapter 7

Neural Networks Results

The results presented below show that neural networks are adept classifiers that are well-suited to generalizing the patterns in the data with which they are presented. This occurs naturally as part of the learning process. In effect, the neural network "filters" the data of its irregularities and noise. This feature is highly desirable since the general trend of the neural network's input-output mapping function is generally more important than the deviations. This is especially true when the deviations are manifested in pattern differences.

7.1 Results of Neural Networks Constructed

The results of the neural networks are presented below. In each case the results indicate the number of correct classifications, both as a fraction and as a percentage. From an analytical perspective, the relative fraction carries more useful information since it indicates the success rate relative to the dataset size. So an error of one misclassification in a dataset of 10 appears (incorrectly) more significant than a single error in a database of 50 when considered as a percentage.

Each neural network is tested on the training data to ensure that it has trained well, and on the test set to ensure that it has generalised well. Generalisability ensures that the neural networks have learnt the underlying discriminatory characteristics of the training dataset rather than the dataset itself. Therefore the test set is able to validate the ability of the neural networks to interpolate between the datapoints of the training dataset. This ability means that the neural networks will perform well on novel (unseen) or noisy data.

7.1.1 Results of the Classification of the Whole Data set into limonoids, triterpenoids, or "other" (NN1)

7.1.1.1 Results of the Classification of the Whole Data set into limonoids, triterpenoids, or "other" (NN1) using 5 ppm binning

The table below shows that the best test results were achieved in **Neural Network:Class37** in which the neural network was trained on both the pure and impure data as opposed to only clean data as in **Neural Network:Class38** and **Neural Network:Class39**. The larger number of compounds in the training set of **Neural Network:Class37** suggested that the more data available, the better the overall performance of the neural network. Secondly, the fact that there were no hidden neurons present in **Neural Network:Class37** implied that the larger the training data set, the more likely it is to see a linearly separable relationship between the inputs.

Secondly, the fact that both pure and impure data was used for training in **Neural Network:Class37** suggested that the neural network was able to generalise better in the recognition of the three classes which is one of its strengths. This is in direct contrast to traditional classification methods that would require spectral data of absolutely pure compounds for optimal training and classification. As the classification in NN1 involved three distinctly different classes, it would appear that purity was not a priority for neural network training in NN1, but rather training set size. Also, the fact that the flavonoid/coumarin representation in the dataset had to be increased so as to make it more representative indicates that neural networks classify best when there are equal representations of all classes.

The results also suggested that neural networks were good at applying learnt knowledge to new sets of test data as shown by the excellent performance of **Neural Network:Class37** in classifying correctly the 16 pure and 16 impure compounds in the unseen test set. Thus, given a good representative training set, neural networks accommodate both unseen pure and impure data exceptionally well.

Table 7.1

Neural Network Name	NN1 Neural Network: Class37	NN1 Neural Network: Class38	NN1 Neural Network: Class39
Training set (seen)	132 pure & impure	53 pure	69 pure
Test set (unseen)	16 pure & 16 impure	16 pure	66 impure
Bin size	5 ppm	5 ppm	5 ppm
Train Set Results	130/132 = 98%	53/53 = 100%	69/69 = 100%
Test Set Results	32/32 = 100%	15/16 = 94%	61/66 = 92%

Sensitivity Analyses

As discussed in Chapter 6, sensitivity analysis reveals which fields need to be changed, by how much and in which direction in order to achieve the desired model output. High sensitivity to particular inputs that are likely to fluctuate around their current values may lead to a questionable output. Overall, the sensitivity analyses for all three networks involving 5 ppm binning, **Neural Network:Class37**, **Neural Network:Class38** and **Neural Network:Class39** were thus very good in that very few input fields were shown to be sensitive. Secondly, as only small perturbations were necessary to make the fields less sensitive, it was concluded that all three networks were stable. **Neural Network:Class37** and **Neural Network:Class39** both had 5 out of a possible 44 sensitive input fields (11% of the total input fields), although **Neural Network:Class39** had only 69 compounds in its training set compared with 132 compounds in the training set of **Neural Network:Class37**. The fact that the number of sensitive input fields did not increase proportionately with an increase in dataset size, was evidence that neural networks improve in performance with bigger and more representative training sets. Further evidence for this was found in **Neural Network:Class37** which had only one more sensitive input field than **Neural Network:Class38** (5 as opposed to 4 (11% as opposed to 9% of the total input fields)) despite having more than double the compounds in the training set (132 as opposed to 53). Clearly, generalisation, which is the strength of neural networks, requires a large, representative training set in order to perform well.

Contribution Analyses

In complex molecules such as those used in this study, carbons representing the same functional groups in different compounds are not found at exactly the same frequency in the spectra. The carbon signals fall within a range of ppm values of which a specific 5 ppm bin may only represent a part. Thus signals representing the same functional groups in different compounds are not restricted to one specific 5 ppm bin, but may be shifted up or downfield into other bins depending on the carbons' different molecular environments.

The results of the contribution analysis of the 5 ppm binned data of the clean data set (**Neural Network:Class38**) showed that the bins between 140-145, 110-115, 80-85 and the 35-40 ppm bins were important in the decision making ability of the neural network to discriminate between limonoids, triterpenoids or "other"(flavonoids/coumarins) given only clean sample data on which to train. Both the bins between 110-115 ppm and 140-145 ppm are important identifiers of limonoids as they represent the furan ring ^{13}C resonances for C-22 (~ 110-112 ppm), C-21 (~140 ppm) and C-23 (~143 ppm) respectively. It was found that the 110-115 ppm bin was particularly important to the classification of the triterpenoids as different to the limonoids and flavonoids/coumarins ("other") with the 140-145 ppm bin following close behind in relevance. It was interesting that the bin between 80 and 85 ppm was chosen by the neural network since the complex prieurianin, evodulone and obacunol type limonoids (C-4 in the opened ring A) as well as the ekebergin type limonoids (C-8,C-14 and C-17 are attached to oxygen atoms forming ether groups within the molecule) resonate in this region. Thus this subset of the limonoids is recognised as belonging to the limonoid set by the neural network. However, this was not an absolute distinction as one dammarane and one glabretal type limonoid also had resonances in this region. Finally, the 35-40 ppm bin which falls within the methyl, methylene and methine region was very important in identification of the flavonoids/coumarins ("other"), as the flavonoid/coumarin class has no resonances in this bin, whereas signals are prevalent in this bin for the triterpenoids, protolimonoids and limonoids.

Table 7.2

Neural Network:Class 38	Important Bin Contributors (in ppm)
triterpenoids	110-115 was more important than 140-145
limonoids	35-40 & 10-15 primarily, but mainly scattered
other	35-40

The results of the contribution analysis of the 5 ppm binned data of the whole data set (**Neural Network:Class37**) however, showed that in addition to the abovementioned bins, the bins between 180-175, 155-160, 105-110 and 35-40 ppm bins were important in the decision making ability of the neural network to discriminate between limonoids, triterpenoids or "other"(flavonoids/coumarins) when trained on the spectra of both pure and impure samples.

The region between 180-175 ppm was also an interesting choice. The region is indicative of carbonyl groups such as acetates and carbomethoxy groups which were predominant in the limonoids, less common in the triterpenoids and only present in three of the flavonoids and not at all present in the coumarins of the data set on which the neural network was trained. Signals indicating the presence of heteroaromatics, aromatics attached to alkyl and polar substituents and polar substituents attached to an unsaturated group within the compound occur in the region between 155 and 160 ppm. Thus this region is particularly important in implying the presence of flavonoids and coumarins. However, this is not an absolute rule as only two coumarins had resonances in this field and one flavonoid had no resonances in this field. To increase the non-linearity of the problem, resonances in this field were present in a number of limonoids as well (α,β-unsaturated ketones and unsaturation between carbon atoms 14 and 15). The region between 105 and 110 ppm cannot be a singly deciding bin as it is again indicative of unsaturation and aromatic nature which is fairly widespread. However, although C-20 of the furan ring resonates at ~109 in some of the limonoids, resonances in this region are consistently present in the flavonoids. But there are no resonances in that bin for the coumarins. All these factors point to the learned complexity of the neural network decision parameters.

Table 7.3

Neural Network:Class 37	Important Bin Contributors (in ppm)
triterpenoids	140-145
limonoids	105-110 was more important than 175-180
other	35-40

Learning Vector Quantisation (LVQ) Results

The results of the three 5 ppm binned neural networks are shown in the table below. Both **Neural Network:Class37** and **Neural Network:Class38** had unseen test sets and the results obtained using *Predict* were slightly better than LVQ. This suggested that neural networks (*Predict*) were good at generalising the pattern recognition of a number of compounds in a training set and then applying this knowledge to new sets of data. In contrast, the test set results for **Neural Network:Class39** using LVQ were considerably better than those using *Predict*. The test set involved was the spectral data of the impure compounds of the pure compounds that were used in the training set. This suggested that LVQ is good at recognising impure data of already known pure data, but rather weak at coping with unseen sets of data.

Table 7.4

	Training set results	Test set results
Neural Network:Class 37	100%	91.6%
Neural Network:Class 38	100%	90.3%
Neural Network:Class 39	100%	97.8%

Back-propagation neural networks results

The results of the three 5 ppm binned neural networks, **Neural Network:Class37**, **Neural Network:Class38** and **Neural Network:Class39** are shown in the table below. It was evident that the back-propagation neural networks achieved the best results when compared with both LVQ and *Predict*. Besides normalisation, no data pre-processing was carried out on the input data in the back propagation neural networks, and thus the neural network structures were far more complex (20-30 hidden neurons) than those in *Predict*.

Table 7.5

	Training set results	Test set results
Neural Network:Class 37	100%	100%
Neural Network:Class 38	100%	100%
Neural Network:Class 39	100%	98.9%

7.1.1.2 Results of Classification of Whole Data set into limonoids, triterpenoids, or "other" (NN1) Using 10 ppm binning

The table below shows that the best test results were achieved **in Neural Network:Class60** in which the neural network was trained on both the pure and impure data as opposed to only clean data as in **Neural Network:Class29** and **Neural Network:Class59**. The trends as discussed under 5 ppm binning above seemed to be applicable here too. However, the overall poorer performance results suggested that 10 ppm binning was too large a bin size, thus obscuring important discriminatory features. This was substantiated by the fact that a linearly separable relationship was suggested (as opposed to a non-linear relationship at 5 ppm) which meant that the 10 ppm averaging had oversimplified the problem.

Table 7.6

Neural Network	NN1	NN1	NN1
Name	**Neural Network:Class60**	**Neural Network:Class29**	**Neural Network:Class59**
Training set (seen)	132 pure & impure	53 pure	69 pure
Test set (unseen)	16 pure & 16 impure	16 pure	66 impure
Bin size	10 ppm	10 ppm	10 ppm
Train Set Results	126/132 = 95%	51/53 = 96%	65/69 = 94%
Test Set Results	30/32 = 94%	14/16 = 88%	56/66 = 85%

Sensitivity Analyses

Overall, the sensitivity analysis for all three networks involving 10 ppm binning, **Neural Network:Class60**, **Neural Network:Class29** and **Neural Network:Class59** were not as good as those involving 5 ppm binning in that more input fields were shown to be sensitive. However, the perturbations that were

necessary to make the fields less sensitive were still of a very small magnitude, showing that all three networks were still fairly stable. **Neural Network:Class60** had 10, **Neural Network:Class29** had 7 and **Neural Network:Class59** had 8 sensitive input fields, which corresponded to 45%, 32% and 36% of the total input fields.

7.1.2 Classification of triterpenoids set into protolimonoids and triterpenoids (NN2(a))

This network was trained to discriminate between the protolimonoids (which included the glabretal-type protolimonoids) and other triterpenoids (which included the dammaranes, masticadienoic acids and plant sterols) which is far more difficult than NN1 as the protolimonoids fall into the general category of triterpenoids and have no distinguishing characteristics. A number of substructural features such as alkenes, alcohols, ethers, carboxylic acids, acetate groups and cyclopropane rings occur not only within each compound, but are frequently common to all of them. Added to this are stereoisomeric mixtures, mixtures of esters and dirty samples.

7.1.2.1 Classification of triterpenoids set into protolimonoids and triterpenoids using 5 ppm binning

As was to be expected, the results when compared with NN1 were poorer for **Neural Network:Class24**, **Neural Network:Class30** and **Neural Network:Class16**, as the decision boundary between classes becomes more indistinct. The table below shows that the best test results were achieved in **Neural Network:Class24** in which the neural network was trained on both the pure and impure data as opposed to only pure data as in **Neural Network:Class30** and **Neural Network:Class16**. The larger number of compounds in the training set of **Neural Network:Class24** suggested that the more data available, the better the overall performance of the neural network. It was difficult, though, to compare the results from **Neural Network:Class24** and **Neural Network:Class 30** directly because the test sets varied so much in size. However, considering the difficulty in separating these two classes, the results still suggested that neural networks were excellent at applying learnt knowledge to new, unseen sets of test data as shown by

the performance of **Neural Network:Class24** in classifying 83% correctly the 6 pure and 6 impure compounds in the unseen test set.

One of the two compounds that were misclassified in the test set of **Neural Network:Class24** was one of the dysoxylic acids which was classified as a triterpenoid rather than a protolimonoid. As already discussed, this was not really an error as the dysoxylic acids are borderline between glabretal-type protolimonoids and masticadienoic acids. Thus, given a good representative training set of data, neural networks can be used to solve fairly complex non-linear problems.

Table 7.7

Neural Network	NN2(a)	NN2(a)	NN2(a)
Name	**Neural Network:Class24**	**Neural Network:Class30**	**Neural Network:Class16**
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure and impure	6 pure	28 impure
Bin size	5 ppm	5 ppm	5 ppm
Train Set Results	46/46 = 100%	24/24 = 100%	27/30 = 90%
Test Set Results	10/12 = 83%	4/6 = 67%	22/28 = 79%

Sensitivity Analyses

Overall, the sensitivity analyses for all three networks involving 5 ppm binning, **Neural Network:Class24**, **Neural Network:Class30** and **Neural Network:Class16** suggested that all three networks were stable and reliable, in that only a few input fields were shown to be sensitive and only small perturbations were necessary to make the fields less sensitive. **Neural Network:Class24** had 7, **Neural Network:Class30** had 5 and **Neural Network:Class16** had 6 sensitive input fields, which corresponded to 16%, 11% and 14% respectively of the total input fields. The fact that **Neural Network:Class30** had only 24 compounds in the training set compared with 46 compounds in the training set of **Neural Network:Class24**, showed that the most stable networks are dependent on the bigger and more representative the training set, rather than purity of compounds contained therein.

Contribution Analyses

The results of the contribution analysis of the 5 ppm binned data of the clean triterpenoid set (**Neural Network:Class30**) showed that the bins between 145-150, 100-105, 95-100, 65-70 and 10-15 ppm bins were important in the decision making ability of the neural network to discriminate between protolimonoids and triterpenoids when trained to recognise spectra of only the clean data. Resonances in the 145-150 ppm bin were particularly indicative of the singlet at C-8 in protolimonoids which had a double bond between carbons 7 and 8 as well as the singlet at C-4 in some of the secA protolimonoids and also the doublet at C-24 where there was a double bond present between carbons 24 and 25 in the uncyclised side chains of the masticadienoic acids and glabretal-type compounds.

The bins between 100-105 ppm and 95-100 ppm were particularly important in conveying the presence of a doublet due to C-21 (~96-102 ppm) when attached to two oxygen atoms i.e.: to a hemiacetal group in a protolimonoid structure. Some of the protolimonoids were present as C-21 epimeric mixtures and thus C-21 would have two representative peaks i.e. pairing. These aforementioned bins, especially 100-105 ppm were, thus very important in the differentiation of the triterpenoids from the protolimonoids. Signals in the bin between 65 and 70 ppm often meant the presence of a doublet at C-24 when in an epoxide group which were present in some of the protolimonoids and glabretal-type protolimonoids but not in the dammaranes, plant sterols and masticadienoic acids.

Table 7.8

Neural Network:Class 30	Important Bin Contributors (in ppm)
triterpenoids	100-105 and 95-100 were more important than 145-150
protolimonoids	randomly spread

The results of the contribution analysis of the 5 ppm binned data of the whole triterpenoid set (**Neural Network:Class24**) showed that in addition to the abovementioned bins, the bins between 130-135, 50-55, 35-40 and 15-20 ppm bins were important in the decision making ability of the neural network to discriminate between protolimonoids and triterpenoids when trained to recognise the spectra of both the pure and impure samples. The resonances in the 130-135 ppm bin

accounted for the singlet at C-25 in the uncyclised side chain in sec A protolimonoids, dammaranes and some of the glabretal-type compounds where there was a double bond between carbons 24 and 25. The remaining bins that were used occur fairly commonly in all of the triterpenoids and protolimonoids as they fall within the methyl, methylene and methine regions.

Table 7.9

Neural Network:Class 24	Important Bin Contributors (in ppm)
triterpenoids	145-150, 100-105
protolimonoids	randomly spread

LVQ results

The results of the three 5 ppm binned neural networks are shown in the table below. All three results were better than those obtained using *Predict*. This suggested that when dealing with compounds which have very small differences as in this case, the generalisation of neural networks (*Predict*) loses a lot of the fine detail on which the compounds could be discriminated.

Table 7.10

	Training set results	Test set results
Neural Network:Class 24	100%	91.6%
Neural Network:Class 30	100%	100%
Neural Network:Class 16	100%	92.5%

Back-propagation neural networks results

The results of the three 5 ppm binned neural networks, **Neural Network:Class24**, **Neural Network:Class30** and **Neural Network:Class16** are shown in the table below. It was evident once again that the back-propagation neural networks achieved the best results when compared with both LVQ and *Predict*. Besides normalisation, no data pre-processing was carried out on the input data in the back propagation neural networks, and thus the neural network structures were far more complex (20 hidden neurons) than those in *Predict*.

Table 7.11

	Training set results	Test set results
Neural Network:Class 24	100%	100%
Neural Network:Class 30	100%	100%
Neural Network:Class 16	100%	100%

7.1.2.2 Classification of triterpenoids set into protolimonoids and triterpenoids Using 10 ppm binning

The table below shows that although good results were obtained on the training sets, the test set results for the unseen data in **Neural Network:Class20** and **Neural Network:Class33** were poor, which suggested that 10 ppm binning was too large a bin size, thus obscuring important discriminatory features.

Table 7.12

Neural Network	NN2(a)	NN2(a)	NN2(a)
Name	**Neural Network:Class20**	**Neural Network:Class33**	**Neural Network:Class12**
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure & impure	6 pure	28 impure
Bin size	10 ppm	10 ppm	10 ppm
Train Set Results	45/46 = 98%	24/24 = 100%	27/30 = 90%
Test Set Results	6/12 = 50%	4/6 = 67%	21/28 = 75%

Sensitivity Analyses

Overall, the sensitivity analyses for all three networks involving 10 ppm binning, **Neural Network:Class20**, **Neural Network:Class33** and **Neural Network:Class12** were not as good as those involving 5 ppm binning in that more input fields were shown to be sensitive. **Neural Network:Class20** had 12 sensitive input fields which corresponded to 55% of the total number of input fields. This indicated a fairly unstable network and the poor test results (50%) reflected this. **Neural Network:Class33** and **Neural Network:Class12** had 8 sensitive input fields, which corresponded to 36% of the total input fields, and hence their test set results were slightly better.

7.1.2.3 Classification of triterpenoids set into protolimonoids and triterpenoids Using 15 ppm binning

There was a slightly improved performance in the results of the 15 ppm binned data of **Neural Network:Class25**, **Neural Network:Class35** and **Neural Network:Class17** shown in the table below compared with the results for the 10 ppm binned data. This was suspicious as the bigger the bin size, the greater the loss of discriminatory information and therefore a poorer overall performance should have occurred. Thus it was suggested that artefacts were formed by the binning process which are not a real reflection of the information content of the data.

Table 7.13

Neural Network Name	NN2(a) **Neural Network:Class25**	NN2(a) **Neural Network:Class35**	NN2(a) **Neural Network:Class17**
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure & impure	6 pure	28 impure
Bin size	15 ppm	15 ppm	15 ppm
Train Set Results	46/46 = 100%	24/24 = 100%	28/30 = 93%
Test Set Results	9/12 = 75%	3/6 = 50%	26/28 = 93%

Sensitivity Analyses

The sensitivity analyses for all three networks involving 15 ppm binning, **Neural Network:Class25**, **Neural Network:Class35** and **Neural Network:Class17** each had 5 sensitive input fields which corresponded to 33% of the total number of input fields. Thus, these sensitivity analyses combined with slightly better test results suggested better networks than those for 10 ppm binning. However, the larger the bin size, the greater the loss in fine detail which would be especially necessary to discriminate between such similar samples. It was concluded that artefacts must have been created which are not a true reflection of the data.

7.1.3 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids (NN2(b))

This network was trained to discriminate between the protolimonoids, the glabretal-type protolimonoids and the rest of the triterpenoids (which included the dammaranes, masticadienoic acids and plant sterols) which, as in NN2(a) is far more difficult than NN1 as the protolimonoids fall into the general category of triterpenoids and have no distinguishing characteristics. Again, a number of substructural features such as alkenes, alcohols, ethers, carboxylic acids, acetate groups and cyclopropane rings occur not only within each compound, but are frequently common to all of them. Added to this are stereoisomeric mixtures, mixtures of esters and dirty samples.

7.1.3.1 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids Using 5 ppm binning

As the discrimination between protolimonoids, glabretal-type protolimonoids and the rest of the triterpenoids is far more non-linear than the discrimination between the three classes in NN1, it was to be expected that the results when compared with NN1 were poorer for **Neural Network:Class27**, **Neural Network:Class31** and **Neural Network:Class52.** However, the results were quite comparable if not slightly better than the results for **Neural Network:Class24**, **Neural Network:Class30** and **Neural Network:Class16** (NN2(a)). As in NN2(a) the table below shows that the best test results were achieved in **Neural Network:Class27** in which the neural network was trained on both the pure and impure data as opposed to only clean data as in **Neural Network:Class31** and **Neural Network:Class52**. The results of **Neural Network:Class27** which had one error in classifying the test set of 6 pure and 6 impure unseen compounds, reinforced the strong capability of neural networks to apply learnt knowledge to new, unseen sets of test data. The two dysoxylic acids were correctly classified as glabretal-type protolimonoids in **Neural Network:Class27** and **Neural Network:Class31** but one was classified as a triterpenoid in **Neural Network:Class52**.

Table 7.14

Neural Network Name	NN2(b) Neural Network:Class27	NN2(b) Neural Network:Class31	NN2(b) Neural Network:Class52
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure & impure	6 pure	28 impure
Bin size	5 ppm	5 ppm	5 ppm
Train Set Results	43/46 = 93%	24/24 = 100%	28/30 = 93%
Test Set Results	11/12 = 92%	5/6 = 83%	23/28 = 82%

Sensitivity Analyses

As only a few input fields were shown to be sensitive and only small perturbations were necessary to make the fields less sensitive, it was concluded that all three networks involving 5 ppm binning, **Neural Network:Class27**, **Neural Network:Class31** and **Neural Network:Class52** were stable and reliable. **Neural Network:Class27** had 8, **Neural Network:Class31** had 6 and **Neural Network:Class52** had 7 sensitive input fields, which corresponded to 18%, 14% and 16% respectively of the total input fields. The fact that **Neural Network:Class31** was the most stable network suggested that when trying to categorise such similar structures as protolimonoids, glabretal-type protolimonoids and triterpenoids, purity of data starts playing a more important role than dataset size.

Contribution Analyses

The results of the contribution analysis of the pure 5 ppm binned data of the triterpenoid set (**Neural Network:Class31**) showed that the bins between 170-175, 145-150, 100-105, 50-55 and 40-45 ppm bins were important in the decision making ability of the neural network to discriminate between protolimonoids, glabretal-type protolimonoids and triterpenoids when trained to discriminate between the spectra of only the clean data.

Resonances in the 170-175 ppm bin (representative of carboxylic acids and esters) were present in the spectra of the masticadienoic acids (-COOH group at C-26 in

the side chain), as well as randomly in the protolimonoids and glabretal-type protolimonoids and the secA protolimonoids.

Resonances in the 145-150 ppm bin were particularly indicative of the singlet at C-8 in protolimonoids which had a double bond between carbons 7 and 8 as well as the singlet at C-4 in some of the secA protolimonoids and also the doublet at C-24 where there was a double bond present between carbons 24 and 25 in the uncyclised side chains of the masticadienoic acids and glabretal-type compounds. The bin between 100-105 ppm were particularly important in conveying the presence of a doublet due to C-21 (~96-102 ppm) when attached to two oxygen atoms i.e.: to a hemiacetal group in a protolimonoid structure. Some of the protolimonoids were present as C-21 epimeric mixtures and thus C-21 would have two representative peaks i.e. pairing.

Resonances in the bin between 50-55 ppm are commonly found in the spectra of the triterpenoids (plant sterols, masticadienoic acids and dammaranes) but on closer examination of the glabretal-type compounds available for this study, it appeared that they were absent in all but two of the glabretals.

Finally the 40-45 ppm bin which falls into the methylene and methine ranges, has resonances present from virtually all three classifications. Due to the fact that this binning process averages the peaks, the C-18 triplet of the glabretal cyclopropane ring in the bin between 13 and 15 ppm was not appropriate as a discriminating factor as most of the compounds had methyl peaks present in this bin.

Table 7.15

Neural Network:Class 31	Important Bin Contributors (in ppm)
triterpenoids	100-105, 170-175, 40-45
glabretal-type protolimonoids	50-55
protolimonoids	170-175

The results of the contribution analysis of the 5 ppm binned data of the whole triterpenoid set (**Neural Network:Class27**) showed that in addition to the abovementioned bins, the bins between 175-180, 125-130, 110-115, 75-80, 30-35, 20-25 and 15-20 ppm bins were important in the decision making ability of the neural network to discriminate between protolimonoids, glabretal-type protolimonoids and triterpenoids when trained to discriminate between the spectra

of both pure and impure data. The region between 180-175 ppm is indicative of carbonyl groups such as those occurring in acetates and carbomethoxy groups which were not present in the plant sterols, the glabretals, the masticadienoic acids or the other protolimonoids. Most of the acetate signals in these compounds occurred between 170-175 ppm except for one dammarane and one secA protolimonoid. With the exception of one of the secA protolimonoids, resonances in the bin between 125 and 130 ppm were absent for the protolimonoids but were present for the majority of the glabretal-type protolimonoids and the triterpenoids (plant sterols, masticadienoic acids and dammaranes with some exceptions). This was because, in this sample set, most of the protolimonoids had cyclised side chains in contrast to the rest of the compounds which had unsaturation in the side chain (often at C-23 or C-25). Thus, this was an important bin on which the protolimonoids could be discriminated. The bin between 75 and 80 ppm which falls within the alcohol and ether range, was fairly common to all these compounds, and the remaining bins that were used (30-35 ppm, 20-25 ppm and 15-20 ppm) occur fairly commonly in all of the triterpenoids and protolimonoids as they fall within the methyl, methylene and methine regions.

Table 7.16

Neural Network:Class 27	Important Bin Contributors (in ppm)
triterpenoids	randomly spread
glabretal-type protolimonoids	30-35
protolimonoids	125-130

LVQ results

The results of the three 5 ppm binned neural networks are shown in the table below. The three test set results were not better than those obtained using *Predict*, and in the case of **Neural Network:Class27**, were worse.

Table 7.17

	Training set results	Test set results
Neural Network:Class 27	93.4%	72.2%
Neural Network:Class 31	95.2%	100%
Neural Network:Class 52	95.8%	95.8%

Back-propagation neural networks results

The results of the three 5 ppm binned neural networks, **Neural Network:Class24**, **Neural Network:Class30** and **Neural Network:Class16** are shown in the table below. It was evident that the back-propagation neural networks achieved the best results when compared with both LVQ and *Predict*. Besides normalisation, no data pre-processing was carried out on the input data in the back propagation neural networks, and thus the neural network structures were far more complex (20 hidden neurons) than those in *Predict*.

Table 7.18

	Training set results	Test set results
Neural Network:Class 27	100%	100%
Neural Network:Class 31	100%	100%
Neural Network:Class 52	100%	100%

7.1.3.2 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using 10 ppm binning

As in NN2(a), the table below showed that although good results were obtained on the training sets, the test set results for the unseen data in **Neural Network:Class26** and **Neural Network:Class32** were poor, although slightly better than in NN2(a) which again suggested that 10 ppm binning was too large a bin size. The two dysoxylic acids were correctly classified as glabretal-type protolimonoids in **Neural Network:Class26** and **Neural Network:Class32** but one was classified as a triterpenoid in **Neural Network:Class54**.

Table 7.19

Neural Network	NN2(b)	NN2(b)	NN2(b)
Name	**Neural Network:Class26**	**Neural Network:Class32**	**Neural Network:Class54**
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure & impure	6 pure	28 impure
Bin size	10 ppm	10 ppm	10 ppm
Train Set Results	45/46 = 98%	24/24 = 100%	28/30 = 93%
Test Set Results	8/12 = 67%	4/6 = 67%	23/28 = 82%

Sensitivity Analyses

The sensitivity analysis for **Neural Network:Class26** showed that it had 9 sensitive input fields, while **Neural Network:Class32** and **Neural Network:Class54** each had 7 which corresponded to 41% and 32% respectively. These were not as good as those involving 5 ppm binning in that more input fields were shown to be sensitive. However, they were slightly better than the 10 ppm binning results in NN2(a) which suggested that the classes chosen in NN2(b) were better than those in NN2(a).

7.1.3.3 Classification of triterpenoids set into protolimonoids, glabretal-type protolimonoids and triterpenoids using 15 ppm binning

The performance in the results of the 15 ppm binned data of **Neural Network:Class28**, **Neural Network:Class34** and **Neural Network:Class53** shown in the table below are comparable with the results for the 10 ppm binned data. This was suspicious as the bigger the bin size, the greater the loss of discriminatory information and therefore a poorer overall performance should have occurred. Thus it was suggested that, as in NN2(a), artefacts were formed which are not a real reflection of the data.

Table 7.20

Neural Network	NN2(b)	NN2(b)	NN2(b)
Name	Neural Network:Class28	Neural Network:Class34	Neural Network:Class53
Training set (seen)	46 pure & impure	24 pure	30 pure
Test set (unseen)	12 pure & impure	6 pure	28 impure
Bin size	15 ppm	15 ppm	15 ppm
Train Set Results	42/46 = 91%	24/24 = 100%	29/30 = 97%
Test Set Results	7/12 = 58%	4/6 = 67%	22/28 = 79%

Sensitivity Analyses

The sensitivity analyses for all three networks involving 15 ppm binning showed that **Neural Network:Class28** and **Neural Network:Class53** each had 8 sensitive input fields which corresponded to 53% of the total number of input fields, while **Neural Network:Class34** had 6 (40%). As it had been concluded that artefacts must have been created in the 15 ppm binned data, these results were not considered further.

7.1.4 Classification of limonoids as individuals (NN3(a))

The neural networks were trained to recognise each of the 21 individual pure pre-identified limonoids within the limonoid data set. The network was then tested on 22 impure limonoids. Again, the whole spectrum was important as the limonoids can vary considerably in all regions since they may have rings A and/or B and /or D opened, extra side-chains, extra acetate groups, saturated and unsaturated ketones, ethers, lactones and formates. Some of the ekebergin type limonoids were mixtures of esters.

7.1.4.1 Classification of limonoids as individuals using 5 ppm and 10 ppm binning

The test set results in both the 5 ppm binned **Neural Network:Class40** and 10 ppm binned **Neural Network:Class13** were very poor, as shown in the table below.

Table 7.21

Neural Network Name	NN3 (a) Neural Network:Class40	NN3 (a) Neural Network:Class13
Training set (seen)	pure 22	21 pure
Test set (unseen)	impure 22	22 impure
Bin size	5 ppm	10 ppm
Train Set Results	22/22 = 100%	22/22 = 100%
Test Set Results	13/22 = 59%	14/22 = 64%

An interesting point to emerge in the training set of the limonoids in **Neural Network:Class40** and **Neural Network:Class13** was that it was originally thought that there were 22 different limonoids in the training set. However, the neural networks consistently classified one of the turraflorin limonoids (compound **IV**) isolated by the author as being the same as one that was isolated by another member of the research team. On closer investigation it was found that the limonoids were the same, although their spectra differed slightly in purity and sample concentration and also due to the fact that the two limonoids were isolated from plant material obtained from different sources and isolated at different times by different researchers in the laboratory. The fact that the neural networks overlooked these minor differences, emphasises their robustness as a classification tool. The two original spectra are shown below:

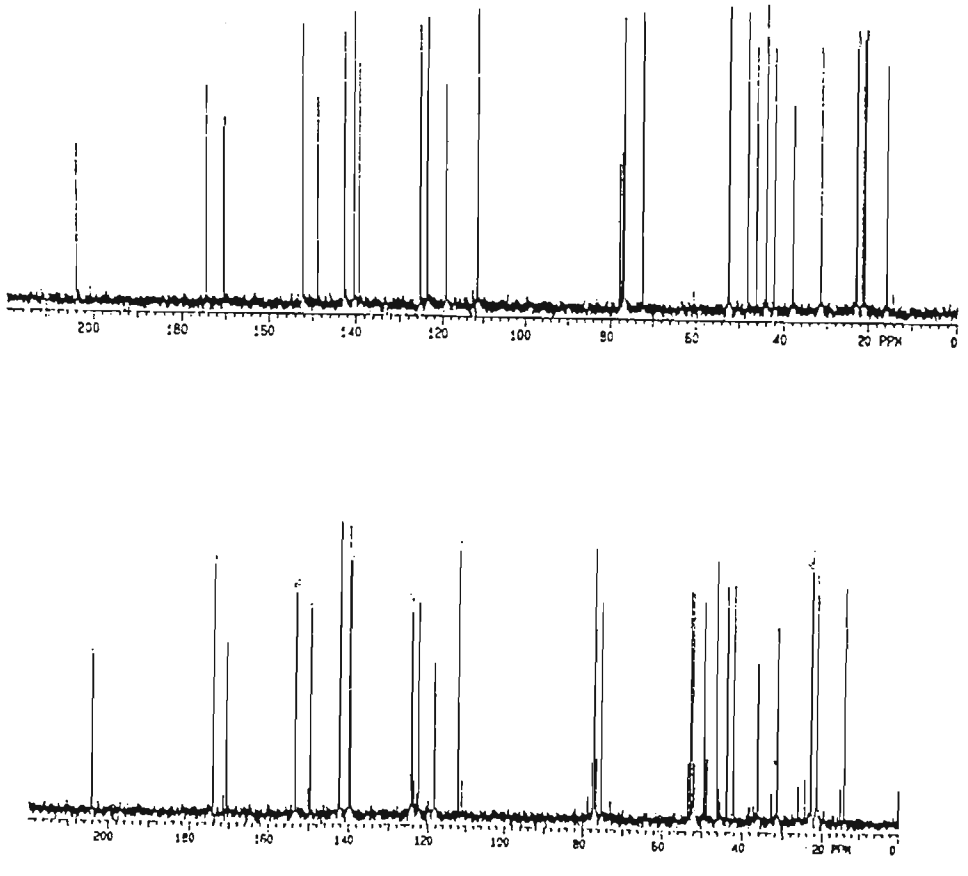


Figure 7.0 a&b Comparison of the original spectra of turraflorin type limonoid.

7.1.5 Classification of limonoids into their 3 Groups using 5 ppm binning (NN3(b))

The table below shows the results obtained when the neural network (**Neural Network:Class56**) was trained to categorize each of the 21 individual pure pre-identified limonoids within the limonoid data set into three groups or classes, according to three criteria:

1. no oxidation i.e.: no carbonyl groups
2. α,β-unsaturation in ring A
3. keto groups at C-9, C-15 or C-3

The limonoids with carbonyl groups in this data set generally had resonances in the fields of 205 ppm to 220 ppm, followed by those with α,β-unsaturation groups which occurred between 200 ppm and 205 ppm, while those simple limonoids with no oxidation groups had no resonances in these regions. This was the only conceivable method able to make use of the unseen test set of limonoids which were a mixture of both pure and impure limonoids, although it was no real challenge to the impure test set. The single misclassification in the 13 limonoids in the unseen test set was an impure mixture.

Table 7.24

Neural Network	NN3(b)
Name	**Neural Network:Class56**
Training set (seen)	pure 22
Test set (unseen)	impure 22
Test set (unseen)	13 pure & impure
Bin size	5 ppm
% Train Set Correct	22/22 = 100%
% Test Set Correct	22/22 = 100%
% Test Set Correct	12/13 = 92%

7.2 Summary

The results of *Predict* showed that better performance was achieved using the 5 ppm binned data rather than the 10 ppm binned data. The slightly improved performance of the 15 ppm binned data over the 10 ppm binned data suggested that this larger bin width was producing artefacts which were not representative of the data, and thus 15 ppm as a binning width size was discarded. Back-propagation neural networks and Learning Vector Quantisation (LVQ) were run on all the 5 ppm binned data, and the excellent results obtained confirmed that the 5 ppm width size was ideal.

In NN1, the 5 ppm binned **Neural Network:Class37** which was trained on 132 pure and impure compounds, and tested on 32 pure and impure unseen compounds achieved the best score with 98% for the training set and 100% for the test set in *Predict*, while the LVQ and Back-propagation scores were both 100% for the training sets and 92% and 100% respectively for the test sets.

In NN2(a), the 5 ppm binned **Neural Network:Class24** which was trained on 46 pure and impure compounds, and tested on 12 pure and impure unseen compounds achieved the best score with 100% for the training set and 83% for the test set in *Predict*, while the LVQ and Back-propagation scores were both 100% for the training sets and 92% and 100% respectively for the test sets.

In NN2(b), the 5 ppm binned **Neural Network:Class27** which was trained on 46 pure and impure compounds, and tested on 12 pure and impure unseen compounds achieved the best score with 93% for the training set and 92% for the test set in *Predict*, while the LVQ and Back-propagation scores were 93% and 95% respectively for the training sets and 72% and 100% respectively for the test sets.

The poor test results in *Predict* for **Neural Network:Class40** in NN3(a) of 60% were surpassed by the excellent LVQ and Back-propagation results of 100% for both the training and test sets.

Finally, **Neural Network:Class56** in NN3(b) scored an excellent 100% on the training set and 92% on an unseen test set using *Predict*.

Chapter 8

Conclusion

The work presented in this thesis was based entirely on data obtained by the author and other members of the Natural Products Research Group from extractives of species belonging to the Meliaceae family. Owing to the uniqueness of the compounds, data from other sources including commercial and research institutions could not be used. The advantages of this specialisation by the Natural Products Research Group are that consistent analytical and elucidation techniques have been established to identify the commonly occurring compounds in these species, as opposed to dealing with a wide variety of compounds from very diverse species belonging to different families.

The computer-based classification of limonoids, protolimonoids and triterpenoids and flavonoids and coumarins performed by the author and presented in this thesis, was unlike the computer-based classification of simpler, synthetic ~C5 compounds or small portions of larger synthetic compounds currently performed by other research groups, in that the compounds were all very complex 3-dimensional structures whose differences ranged from a single functional group (i.e.: an hydroxy group replaced by an acetate group) to major structural changes.

Because the decision rules used by the researchers in the analysis of the NMR spectra of limonoids, protolimonoids, triterpenoids and flavonoids and coumarins are, to some extent, based on experience and knowledge of the biosynthesis of such compounds, algorithmic procedures are limited. Furthermore, expert systems which are a natural option for building classification schemes, are restricted by the tedious and difficult task of extracting and formalising operational expert knowledge based on existing human expertise. Artificial neural networks are good at coping with large amounts of data, with many complex interactions and with uncertain relationships between the data. The layered feedforward networks used in this research undergo a dynamic supervised learning process which did not require

explicit formulation of the relationships that may have existed between the data used as input and the output values.

Identifying substructural features by looking at particular regions in the ^{13}C NMR spectrum and combining the substructures into a larger structure was shown to pose difficulties to the less-experienced as was the case with the incorrect structural assignment of dysobinin (Chapter 3) when only isolated chemical shift positions were taken into account. Secondly, variations in ^{13}C NMR resonances were not confined to any specific regions in the ^{13}C NMR spectrum as limonoids have very diversified structures ranging from all of the rings in the nucleus intact to various permutations of ring A, B, C or D openings. Frequently, extra side chains are present which, due to the 3-dimensional stereochemistry of such large compounds, tend to affect a few regions simultaneously but not consistently. Thus, input vectors for the neural networks based solely on chemical shift positions had to be avoided.

Taking the first 10 strongest peaks also posed a problem as simpler natural products containing a furan ring would not be discriminated against. One possible option considered by the author was to divide the spectrum into chosen regions and peak pick or use small p.p.m. bin widths in those regions and combine all the outputs from those regions into an input vector for the neural networks. This technique may prove successful as an absolute elucidation technique when dealing with absolutely pure synthetic samples. However, in this natural product research, where at times the spectra of impure samples ranged from having approximately 70 extra peaks (limonoids, protolimonoids) to approximately 5 peaks that could not be distinguished above noise levels (coumarins), this technique was unsuitable. Rather than segmenting the problem, the binning technique which was used allowed the entire ^{13}C NMR spectrum "fingerprint" to be used as the data input vector for each compound. Hence, by limiting the amount of human intervention, the capabilities of neural networks were extensively exercised to find complex and obscure relationships between data that aided and improved classification.

Further support of the binning technique came from the fact that it is common in natural product isolation to obtain such small quantities of mixtures that they are inseparable. In this research, one of the glabretal-type protolimonoids and

melianone were both isolated as inseparable stereoisomeric mixtures. Similarly, some of the glabretal-type protolimonoids were isolated as mixtures of cinnamoyl and benzoyl esters while some of the Ekebergin limonoids were isolated as mixed esters of nicotinate and 2-methylbutyrate and where the two isomers in the mixture were in 1:1 proportion, all peaks were included in the data of the pure compound. This resulted in peak duplication, that is for each carbon atom in the compound, two peaks would be present in the ^{13}C NMR spectrum which would be a source of confusion in a peak picking classification technique. Furthermore, due to peak superposition, the number of carbons in a particular compound does not correlate to the number of peaks appearing in the ^{13}C NMR spectrum.

The results of this investigation showed that, using the binning technique, neural networks are extremely proficient at classifying complex natural products. In most of the artificial neural network applications in chemistry to date, the neural networks have been used principally for inferring the presence or absence of substructures from spectra. However, in contrast to picking only specific peaks or selective data inputs, the results of this research showed that the interpretation of the entire spectral range achieved by peak intensity averaging into bins 5, 10 or 15 p.p.m. wide, was a robust technique. For example, it captured the whole spectrum while retaining enough detail for differentiation, and presented it to the neural network as a vector of consistent dimension. These features of the binning technique enhanced the suitability of neural networks to spectroscopic analysis even though spectroscopic signals of the particular "natural products" in this research are noisy and complex and the recognition of specific patterns therein is not straightforward. Excellent results, which showed that the neural networks are highly tolerant to both missing and noisy data, were achieved using both 5 p.p.m. and 10 p.p.m. binning for the following networks in the hybrid neural network architecture:

NN1-the discrimination between limonoids, triterpenoids (protolimonoids, plant sterols, dammaranes and masticadienoic acids) and "other"(flavonoids/coumarins)

NN2(a)-the discrimination between the protolimonoids (including the glabretal-type protolimonoids) and the rest of the triterpenoids (plant sterols, dammaranes and masticadienoic acids)

NN2(b)-the discrimination between the protolimonoids, glabretal-type protolimonoids and the rest of the triterpenoids (plant sterols, dammaranes and masticadienoic acids)

NN3(a)-the discrimination between each individual limonoid in the limonoid data set.

A major advantage in the establishment of a database created for a neural network classification system, was that real data could be used rather than empirical data from books. For example, the limonoid training set (NN3 (a)) had two of the same Turraflorin limonoids which were isolated from the same species, *T. floribunda* by different researchers on separate occasions, and although the compounds were identical, their ^{13}C NMR spectra showed different impurities as is to be expected in natural product isolation. The neural networks however, correctly assigned both these spectra to the same compound. The spectra were compared in Chapter 7.

The superiority of the 5 p.p.m. over the 10 p.p.m. and 15 p.p.m. spectral bin widths suggested that the 5 p.p.m. binned ^{13}C NMR spectrum was sufficiently generalised to cope with stereoisomeric and ester mixtures as well as impure compounds, while at the same time retaining features which made each spectrum distinguishable. Furthermore, the results of the neural networks trained on the 5 p.p.m data were consistently good, whereas the 10 p.p.m. and 15 p.p.m. data produced widely varying results. This stemmed from the fact that the wider 10 p.p.m. and 15 p.p.m. bins appear to have produced averaging and boundary artefacts which may randomly enhance or impair neural network learning. The artefacts arise from averaging which reduces the significance of any one significant peak, from wide bins which cluster too many peaks into one bin, and from boundaries on wide bins which exclude peaks > 0.5 p.p.m. from the boundary edges and centre bin averages far from the original contributory peaks. These artefacts can be seen by comparing the binned spectra in Chapter 6 (6.2).

In contrast, the 5 p.p.m. binning produced far fewer artefacts so that the networks learnt predominantly on the data and not on the artefacts. This was shown clearly in **Neural Network:Class37** and **Neural Network:Class39** of NN1 which used

consistent data transformations (indicating consistent data) in contrast to **Neural Network:Class59** and **Neural Network:Class60** of NN1 which used very different data transformations (indicating data plus artefacts). This was also evident in **Neural Network:Class24** and **Neural Network:Class30** of NN2(a) which also had consistent data transformations as opposed to **Neural Network:Class20** and **Neural Network:Class33,** and **Neural Network:Class25** and **Neural Network:Class35** of NN2(a).

The results of NN1, the limonoid, triterpenoid and "other" classifier, showed the superiority of the 5 p.p.m. binning technique. The neural networks, **Neural Network:Class39** and **Neural Network:Class37**, demonstrate the ability to correctly and consistently identify unseen pure and impure samples. The robustness of these neural networks is seen in their high success rate with the impure data. This implies that noisy samples or spectra obtained from different sources by different processes can be correctly classified with an overall confidence of greater than 92%. The results of NN1 show that neural networks can perform very well on large datasets of complex compounds with varying degrees of impurity.

The performance of neural network NN2(a) identifying the protolimonoids as different from the rest of the triterpenoids is slightly poorer than the rest of the neural networks presented. However, the results (79% accuracy on unseen impure data) are still significant when considering the difficulty of differentiating protolimonoids which are a type of triterpenoid. Again, the neural network trained on the 5 p.p.m. binned data was superior to that using the 10 p.p.m. data.

The results of the performance of neural network NN2(b) which discriminated between the normal protolimonoids, the glabretal-type protolimonoids and the rest of the triterpenoids were markedly better than NN2(a) which suggested that this class division into three groupings was more suitable than just the two groupings of NN2(a). The results (92% accuracy on unseen pure and impure data) are excellent when considering the difficulty of differentiating between the three classes of triterpenoid. Again, the neural network trained on the 5 p.p.m. binned data was superior to that using the 10 p.p.m. data.

The appropriateness of the 5 p.p.m. bin width was reflected in the excellent results obtained using back-propagation and LVQ neural networks. Back-propagation neural networks achieved almost 100% success for each 5 p.p.m. binned neural network trained in NN1, NN2(a) and NN2(b).

Finally, NN3(a) that recognises each of the individual limonoids proved the superiority of Kohonen's self-organising Learning Vector Quantisation (LVQ) for both the 5 p.p.m. and 10 p.p.m. binned **Neural Network:Class40 and Neural Network:Class13** respectively over the multi-layered feedforward neural networks of *Predict*. Unlike the application of NN1, NN2(a) and NN2(b) where there were several examples of each class that the network had to learn, NN3(a) had to learn that each individual limonoid was a separate class. With NN1, NN2(a) and NN2(b), the networks learn regions which relate to each class, but with NN3(a), each class is represented by a single data record without multiple examples. The LVQ method is more suited to data of this sort than other types of networks. In the case of NN3(a), only one output is activated in response to the input data, making the LVQ networks in NN3(a) winner-take-all networks. The "winner" is the output which best matches the desired input. Thus the LVQ network of **Neural Network:Class40** is a robust similarity detector which improves performance to 100% accuracy compared with around 60% for non-LVQ networks. Once again, the 5 p.p.m. data produces consistently better results.

Finally, NN3(b) showed that excellent results were achieved on both pure and impure unseen test sets when grouping the limonoid dataset into three different classes and using 5 p.p.m. binning.

The results of the sensitivity analyses showed that the 5 p.p.m. binned neural networks in NN1 and in NN2(a) and NN2(b) had the least number of sensitive fields out of the total number of input fields, when compared with both the 10 and 15 p.p.m. binned neural networks. Thus it was concluded that the 5 p.p.m. binned neural networks, being the most stable, were most suitable for the training and testing of the data presented in this thesis.

Overall, the contribution analyses revealed that there was much overlapping of important bin contributors within each neural network, which emphasised the non-

linearity of the decision process. However, it must be emphasised that these bin contributors are not absolute rules for the classification of limonoids, protolimonoids, glabretal-type protolimonoids, triterpenoids and flavonoid/coumarins, but are rather representative of the specific training data which have been used to train these neural networks and, as a result, the bin importance would change with added data. Thus neural networks are a dynamic solution as they change according to the data set.

Apart from the application to the extractives of Meliaceae, neural networks show that they can be extended to other sets of extractives, and provide a robust method for the classification of pre-identified natural products. For limonoid researchers, this technique has the potential to ultimately reduce the level of human expertise required as well as the time spent on analysing and interpreting the spectra of pre-identified limonoids. The fact that samples need not be 100% pure (i.e. extraneous peaks, stereoisomeric mixtures and mixtures of esters) means a significant reduction in laboratory chromatographic separation time.

References

Adams, M.J. (1995). Chemometrics in Analytical Spectroscopy. The Royal Society of Chemistry. Cambridge.

Akerman, L-A. (1991). *MSc. Thesis*, University of Natal., Durban, South Africa.

Akerman, L-A., Mulholland, D.A. and Nair, J.J. (1992). New Limonoids from the seed of *Turraea floribunda*. Frank Warren National Organic Chemistry Conference in Gordon's Bay, South Africa.

Akinniyi, J.A., Connolly, J.D., Mulholland, D.A., Rycroft, D.S. and Taylor, D.A. (1986). Limonoid Extractives from *Turraea floribunda* and *Turraea obtusifolia*. *Phytochem.*. **25**, 2187-2189.

Akisanya, A., Bevan, C.W.L., Powell, J.W., Halsall, T.G. and Taylor, D.A.H. (1961). *J. Chem. Soc.* 3705

Akisanya, A., Bevan, C.W.L., Hirst, J., Halsall, T.G. and Taylor, D.A.H. (1960). *J. Chem. Soc.* 3827.

Alam, M.K., Stanton, S.L. and Hebner, G.A. (1994). Near-Infrared Spectroscopy and Neural Networks for Resin Identification. *Spectroscopy*, 30-40.

Ankers, L.S. and Jurs, P.C. (1992). Prediction of ^{13}C NMR Chemical Shifts by using Artificial Neural Networks. *Anal. Chem.* **10**, 1157-1164.

Ansell, S.M. (1986). *MSc. Thesis*, University of Natal, Durban, South Africa.

Arigoni, D., Barton, D.H.R., Bernasioni, R., Djerassi, C., Mills, J.S.and Wolff, R.E. (1960). *J. Chem. Soc.* 1900.

Arigoni, D., Barton, D.H.R., Caglioti, L., Corey, E.J., Dev, S, Ferrini, P.G., Glazier, E.R., Jeger, O., Melera, A., Pradham, S.K., Schaffner, F., Sternhell,S., Templeton, J.F. and Tobinaga, S. (1960). *Experientia.* **16(41)**.

Bevan, C.W.L., Ekong, D.E.U. and Taylor, D.A.H. (1965). *Nature*. **206**, 1323.

Bevan, C.W.L., Ekong, D.E.U., Halsall, T.G. and Toft, P.(1967). *J. Chem. Soc.(C)*. 820.

Bridle, J.S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition in Neurocomputing. NATO ASI Series. F68.

Buchanan, J.G.St.C and Halsall, T.G. (1969). *J. Chem. Soc., Chem. Commun*. 243.

Buchanan, J.G.St.C and Halsall, T.G. (1969). *J. Chem. Soc., Chem. Commun*. 1493.

Carpenter, G.A. and Grossberg, S. (1987a). A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Computer Vision, Graphics and Image Processing*. **37**, 54-115.

Carpenter, G.A. and Grossberg, S. (1987b). ART2: Self-Organization of Stable Category Recognition Codes for Analog Input Patterns. *Applied Optics*. 4919-1930.

Carpenter, G.A. and Grossberg, S. (1988). The ART of Adaptive Pattern Recognition by a Self-Organizing Neural Network. *Computer*. 77-88.

Chatterjee, S. (1991). Regression Analysis by Example. John Wiley & Sons.

Cichocki, A. and Unbehauen, R. (1993) Neural Networks for Optimization and Signal Processing. John Wiley and Sons.

Clouser, D.L. and Jurs, P.C. (1996). Simulation of ^{13}C NMR spectra of Ribonucleosides using linear regression analysis and Neural Networks. *J. Chem. Inf. Comput*. Sci. **36**, 168-172.

Connolly, J.D., Labbe, C., Rycroft, D.S. and Taylor, D.A.H. (1959). Tetranortriterpenoids and related compounds. Part 22. New apotirucallol derivatives and tetranortriterpenoids from the wood and seed of *Chisocheton paniculatis* (Meliaceae). *J. Chem. Soc. Perkin Trans.* **I**, 2959.

Connolly, J.D., Okorie, D.A., de Wit, L.D. and Taylor, D.A.H. (1979). *J.Chem Soc. Chem. Commun.* 909.

Connolly, J.D., Overton, K.H. and Polonsky, J. (1970). *Prog. in Phytochem.* **2**, 385.

Cotterrell, G.P., Halsall, T.G. and Wriglesworth, M.J. (1967). *Chem. Commun.* 1121.

Cotterrell, G.P., Halsall, T.G., Wriglesworth, M.J., Djerassi, D., Mills, J.S. and Wolff, R.E. (1970). *J. Chem. Soc.(C).* 1503.

Curry, B., and Rumelhart, D.E. (1990). MSnet: A neural network that classifies Mass Spectra. *Tetrahedron Comput. Methodol.* **3**. 213-237.

Doucet, J.P., Panaye, A., Feuilleaubois, E. and Ladd, P. (1993). Neural Networks and ^{13}C NMR Chemical Shift Prediction. *J. Chem. Inf. Comput. Sci.* **33(4)**, 320-324.

Dyer, R.A. The genera of South African flowering plants. (1975) 1. Dicotyledons. Pretoria: Department of Agricultural Technical Services. 1975.

Ehrlich, J. and Eymard, R. (1993). An Efficient Implementation of Conjugate Gradient Method for Multilayer Neural Networks in Proceedings of Neural Nimes. EC2.

Fahlman, S.E. (1988). An Empirical Study of Learning Speed in Back-Propagation Networks. Technical Report CMU-CS-88-162, Carnegie-Mellon University.

Fahlmann, S.E. and Lebiere, C. (1988) The Cascade-Correlation Learning Architecture in Advances in Neural Information Processing Systems 2. Morgan Kaufmann.

Fang, S.Y., He, Z.S. and Fan, G.J. (1996). Triterpenoids from *Adina rubella. J. Natl. Prod, ACS.* **59(3)**, 304-307.

Fessenden, R.J. and Gyorgyi, L. (1991). Identifying Functional Groups in IR Spectra Using an Artificial Neural Network. *J. Chem. Soc. Perkin Trans.* **2**, 1755-1762.

Fraser, L-A. and Mulholland, D.A. (1993). Limonoids from the seed of *Turraea obtusifolia.* International conference on Botanical diversity, Cape Town, South Africa, and listed as a contributor to the book, "Botanical Diversity in Southern Africa", edited by B.J.Huntley, Pretoria,1994.

Fraser, L-A. and Mulholland, D.A. (1995). A new limonoid from the seed of *Turraea obtusifolia.* Frank Warren National Organic Chemistry Conference in Bloemfontein, South Africa.

Fraser, L-A., Mulholland D.A. and Fraser, D.D. (1997). Classification of Limonoids and Protolimonoids using Neural Networks. *Phytochem Anal. In press.*

Fraser, L-A., Mulholland, D.A. and Nair J.J. (1992). A New Limonoid from the seed of *Turraea floribunda. Planta medica, J. Med. and Plant Res.* **58**, A575-A724.

Fraser, L-A., Mulholland, D.A. and Nair, J.J. (1994). Limonoids from the seed of *Turraea floribunda. Phytochem.* **35(2)**, 455-458.

Fraser, L-A., Mulholland, D.A. and Taylor, D.A.H. (1995). The chemotaxonomic significance of the limonoid, nymania-1, in *Turraea obtusifolia. S. Afr. J. Bot.* **61(6)**, 281-282.

Gullo, V.P., Miura, I., Nakanishi, K., Cameron, A.F., Connolly, J.D., Duncanson, F.D., Harding, A.E., McCrindle, R. and Taylor, D.A.H. (1975). Structure of prieurianin, a complex tetranortriterpenoid; nuclear magnetic resonance analysis at non-ambient temperatures and x-ray structure determination. *J. Chem. Soc. Chem. Commun.* 345.

Halsall, T.G. and Trobe, J.A. (1975). The structure of three new meliacins isolated from *Khaya anthotheca* heartwood. *J. Chem. Soc. Perkin Trans.* **I**, 1758.

Hasegawa, S, Herman, Z and Ou, P. (1986). *Phytochem.* **25**, 2783.

Haykin, S. (1994). Neural Networks – A comprehensive foundation. MacMillan. New York.

Heller, W. and Forkmann, G. The Flavonoids, Advances in Research. (editor, J.B. Harborne). Chapman and Hall Ltd. London, New York.

Iourine, I. (1996). *PhD. Thesis*, University of Natal., Durban, South Africa.

Ivanciuc, O., Rabine, J.P., Cabrol-Bass, D., Panaye, A. and Doucet, J.P.(1996, in press). ^{13}C NMR Chemical shift prediction of sp2 carbon atoms in acyclic alkenes using neural networks. *J. Chem. Inf. Comput. Sci.*

Jacobs, R.A. (1988). Increased Rates of Learning through Learning Rate Adaptation. *Neural Networks.* **1**, 295-307.

Jibodu, K.O., Ohochuku, N.S. and Taylor, D.A.H. (1970). Chemical shift of the tertiary methyl groups in the nuclear magnetic resonance spectra of some limonoids. *J. Chem. Soc. C.* **II**, 2396.

Jussieu, A.L. De. (1789). *Genera Plantarum.* 263-266.

Keller, P.E., Kangas, L.J., Troyer, G.L., Hashem, S and Kouzes, R.T. (1995). Nuclear Spectral Analysis via Artificial Neural Networks for Waste Handling. *IEEE Transactions on Nuclear Science.* **42**, 709-715.

Kemp, W. (1991) Organic Spectroscopy. Macmillan. London.

Khalid, S.A. (1989). *J. Natl. Prod.* **52**, 922.

Klimasauskas, C. (1992) Neural Network Development Methodology. NeuralWare.

Kohonen, T. (1988). Statistical Pattern Recognition with Neural Networks: Benchmark Studies. *Proc. of the 2nd Annual IEEE International Conference on Neural Networks*. **1**.

Kononenko, I. and Bratko, I. (1991) Information-Based Evaluation Criterion for Classifier's Performance in Machine Learning. 6, pp. 67-80. Kluwer Academic.

Kosela, S., Yulizar, Y., Chairul, C., Tori, M., and Asakawa, Y. (1995). Secomultiflorane-type triterpenoid acids from stem bark of *Sandoricum koetjape*. *Phytochem.* **38(3)**, 691-694.

Kosko, B. (1987). Bidirectional Associative Memories. *IEEE Trans. on Systems, Man, and Cybernetics*.

Koza, J.R. (1993) Genetic Programming. MIT Press.

Kraus, W., Grimminger, W. and Sawitzki, G. (1978). *Angew. Chem.* **17**, 452.

Kubo, I. and Klocke, J.A. *INRA Colloque.* (1981). 7, 117 in Taylor, D.A.H. The Chemistry of the Limonoids from Meliaceae. Springer-Verlag. New York. 1984.

Kubota, K., Fukamiya, N., Hamada, T., Okana, M., Tagahara, K., and Lee, K-S. (1996). Two New Quassinoids, Ailantinols A and B, and Related Compounds from *Ailanthus altissima*." *J. Natl. Prod., ACS.* **59**(7), 683-686.

Kvasnicka, V. (1991). An Application of Neural Networks in Chemistry. Prediction of ^{13}C NMR Chemical Shifts. *J. Math. Chem.* **6**, 63-76.

Kvasnicka, V., Sklenak, S. and Pospichal, J. (1992(a)). Application of Recurrent Neural Networks in Chemistry. Prediction and classification of ^{13}C NMR Chemical Shifts in a series of monosubstituted benzenes. *J. Chem. Inf. Comput. Sci.* **32(6)**, 742-747.

Kvasnicka, V., Sklenak, S. and Pospichal, J. (1992(b)). Application of Neural Networks with feedback connections in chemistry: Prediction of ^{13}C NMR chemical shifts in a series of monosubstituted benzenes. *J. Molec. Struct. (Theochem.)*. **277**, 87-107.

Lam, L.K.T., Hasegawa, S. (1987). *Nutrition and Cancer.* **12**, 43.

Lam, L.K.T., Li, Y. and Hasegawa, S. (1989). *J. Amer. Chem. Soc.* **37**, 878.

Lerner, J.M. and Lu, T. (1993). Practical Neural Networks Aid Spectroscopic Analysis. *Photonic Spectra*, pp. 93-98.

Lohninger,H. and Stancl, F. (1992). Comparing the performance of neural networks to well-established methods of multivariate data analysis: the classification of mass spectral data. *Fresenius' J. Anal Chem.* **344**, 186-189.

Lukacova, V., Polonsky, J., Moretti, C., Pettit, G.R. and Schmidt, J.M. (1982). Isolation and structure of 14b,15b epoxy prieurianin from the South American tree *Guarea guidona*. *J. Natl. Prod.* **45**, 288.

Maclachlan, L. (1981). *MSc. Thesis*, University of Natal, Durban, South Africa.

Maclachlan, L.K. and Taylor, D.A.H. (1982). Limonoids from *Nymania capensis*. *Phytochem.* **21**, 1701-1703.

Maren, A.J., Harston, C.T. and Pap, R.M. (1990). Handbook of Neural Computing Applications. Academic Press, Inc. San Diego. USA.

Meyer, M. and Wiegelt, T. (1992). Interpretation of Infrared Spectra by Artificial Neural Networks. *Anal. Chim. Acta.* **265**, 183-190.

Miller, E.G., Fanous, R., Riveras-Hidago, F., Binnie, W.H., Hasegawa, S and Lam, LKT. (1989). *Carcinogenesis.* **10**, 1535.

Minai, A.A. and Williams, R.D. (1990). Acceleration of Back-propagation Learning through Learning Rate and Momentum Adaptation. *International Joint Conference on Neural Networks.* **I,** 676-679.

Miyashita, Y., Yoshida, H., Yaegashi, O., Kimura, T., Nishiyama, H. and Sasaki, S. (1994). Non-Linear Modelling of ^{13}C NMR Chemical shift data using Artificial Neural Networks and Partial Least Squares Method. *J. Mol. Struct. (Theochem.).* **117**, 241-245.

Monkhe, T.V. (1991). *MSc Thesis*, University of Natal, Durban, South Africa.

Mootoo, B.S., Ramsewak, R., Khan, A., Tinto, W.F., Reynolds, W.F., McLean, S. and Yu, M. (1996). Tetranortriterpenoids from *Ruagea glabra. J. Natl. Prod, ACS.* **59(5)**, 544-547.

Moss, G.P. (1966). *Planta Med.* (*suppl.*). 86.

Mulholland, D.A. and Monkhe, T.V. (1993). Two glabretal-type triterpenoids from the heartwood of *Aglaia ferruginaea. Phytochem.* **34(2)**, 579-580.

Mulholland, D.A. and Nair, J.J. (1994). Triterpenoids from *Dysoxylum pettigrewianum.. Phytochem.* **37(5)**, 1409-1411.

Mulholland, D.A., Nair, J.J. and Taylor, D.A.H. (1994). Dammaranes from *Astrotrichilia asterotricha. Phytochem.* **35(2)**, 542-544.

Mulholland, D.A., Nair, J.J. and Taylor, D.A.H. (1996). Astrotrichilin, a limonoid from *Astrotrichilia asterotricha. Phytochem.* **42(4)**, 1239-1241.

Mulholland, D.A., Nair, J.J. and Taylor, D.A.H. (1996). Glabretal triterpenoids from *Dysoxylum muelleri. Phytochem.* **42(6)**, 1667-1671.

Mulholland, D.A., Osborne, R., Roberts, S.L. and Taylor, D.A.H. (1994). Limonoids and triterpenoid acids from the bark of *Entandrophragma delevoyi*. *Phytochem.* **37(5)**, 1417-1420.

Munk, M.E., Madison, M.S. and Robb, E.W. (1991). Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta (Wien).* **2**, 505-514.

Munk, M.E., Madison, M.S. and Robb, E.W. (1996). The Neural Network as a tool for Multispectral Interpretation. *J. Chem. Inf. Comput. Sci.* **36**, 231-239.

Naidoo, N. (1997). *PhD. Thesis*, University of Natal, Durban, South Africa.

Nair, J.J. (1994).*MSc. Thesis*, University of Natal., Durban, South Africa.

Neural Computing. (1993). NeuralWare Inc., Pittsburgh, PA (no listed author).

Ohochuku, N.S. and Taylor, D.A.H. (1969). Chemical shift of the tertiary methyl groups in the nuclear magnetic resonance spectra of some limonoids. *J. Chem. Soc. C*. 864.

Okorie, D.A. and Taylor D.A.H. (1977). Triterpenes from the seed of *Entandrophragma* species. *Phytochem.* **16**, 2029-2030.

Olmos,P., Diaz, J.C., Perez, J.M., Garcia-Belmonte, G., Gomez,P and Rodellar,V. (1992). Application of neural network techniques in gamma spectroscopy. *Nuclear Instruments and Methods in Physics Research.* **A312**, 167-173.

Panaye, A., Doucet, J.P., Fan, B.T., Feuilleaubois, E., Rahali el Azzouzi, S. (1994). Artificial Neural Network simulation of ^{13}C NMR shifts for methyl substituted cyclohexanes. *Chemometrics and Intel. Lab Systems.* **24**, 129-135.

Parker, D.B. (1985). Learning Logic. Technical Report TR-47. Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA,USA.

Pennington and Styles. (1975). *Blumea.* **22**, 460.

Plutowski, M. and White, H. (1992) Selecting Concise Training Sets from Clean Data. University of California Institute for Neural Computation and Department of Economics.

Puskorius, G. and Feldkamp, L. (1991) Decoupled Extended Kalman Filter Training of Feedforward Layered Networks in Proceedings of the IJCNN, IEEE.

Ricard, D., Cachet, C., Cabrol-Bass, D. and Forrest, T.P. (1993). Neural Network Approach to Structural Feature Recognition from Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **33**, 202-210.

Robb, E.W. and Munk, M.E. (1990). A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta (Wien).* **1**, 131-155.

Roberts, S.L. (1994). *MSc. Thesis*, University of Natal, Durban, South Africa.

Rumelhart, D.E. , McClelland, J.L. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. *MIT Press*, Cambridge MA., USA.

Samad, T. (1989). Back–Propagation Extensions. Honeywell SSDC Technical Report, 1000 Boone Ave.N., Golden Valley, MN 55427, USA.

Silverstein, R.M., Bassler, G.C. and Morrill, T.C. (1981). Spectrometric Identification of Organic Compounds. John Wiley & Sons.

Singh, S., Garg, H.S. and Khanna, N.M. (1978). Dysobinin, a new tetranortriterpene from *Dysoxylum binectariferum*. *Phytochem.* **15**, 2001.

Smits, J.R.M., Melssen, W.J., Derksen, M.W.J. and Kateman, G. (1993). Drift correction for pattern classification with neural networks. *Anal.Chim Acta.* **284**, 91-105.

Sonder, O.W. (1860). Meliaceae. In: W.H. Harvey & O.W. Sonder. Flora capensis. **1**. 244-248.

Specht, D.F. (1988). Probabilistic Neural Networks for Classification, Mapping and Associative Memory. *Proc. International Conference on Neural Networks.*

Specht, D.F. (1990). Probabilistic Neural Networks. Neural Networks.

Svozil, D., Pospichal, J. and Kvasnicka, V. (1995). Neural Network prediction of ^{13}C NMR chemical shifts of Alkanes. *J. Chem. Inf. Comput. Sci.* **35**, 924-928.

Takeya, K., Kobata, H., Ozeki, A., Morita, H. and Itokawa, H. (1997). A New Quassinoid from *Ailanthus vilmoriniana*. *J. Natl. Prod., ACS.* **60**(6), 642-644.

Tanabe, K., Tamura, T. and Uesaka, H. (1992). Neural Network System for the Identification of Infrared Spectra. *Appl. Spectrosc.* **46 (5)**, 807-810.

Taylor, A. (1983). *MSc. Thesis*, University of Natal, Durban, South Africa.

Taylor, D.A.H. (1984). The Chemistry of the Limonoids from Meliaceae. *Progress in the Chemistry of Organic Natural Products*. **45**. Springer Verlag.

Thomson, J.U. and Meyer, B. (1989). Pattern Recognition of the ^{1}H NMR spectra of Sugar Alditols Using Neural Networks. *J. Magn. Reson.* **84**. 212-217.

Tukey, J.W. (1977) Exploratory Data Analysis. Addison Wesley.

Ventenat, E.P. (1799). Tableau du Regne Vegetal selon la methode de Jussieu. **3**, 150-166.

Weigel, U.-M. and Herges, R. (1992). Automatic Interpretation of Infrared Spectra: Recognition of Aromatic Substitution Patterns using Neural Networks. *J. Chem. Inf. Comput. Sci.* **32**, 723-731.

Werbos, P.J., (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioural Sciences", *PhD thesis*, Applied Math., Harvard University, USA.

Werbos, P.J., (1988). Backpropagation:Past and Future. *Proc. IEEE Intl. Conf. On Neural Networks*, **I**, 343-353, New York, USA.

Widrow, B. and Stearns, S.D. (1985) Adaptive Signal Processing. Prentice-Hall.

Widrow, B. (1960). An Adaptive Adeline Neuron using chemical Memistors. Technical Report Number 1553-2. Stanford Electronics Laboratories.

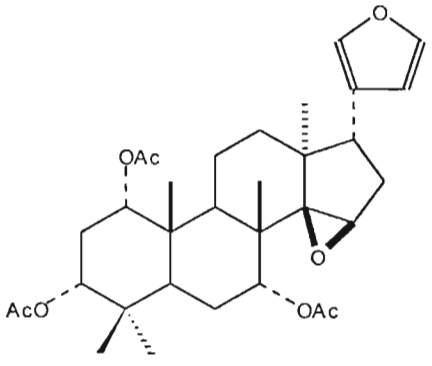
Wythoff, B.J., Levine, S.P. and Tomellini, S.A. (1990). Spectral Peak Verification and Recognition Using a Multilayered Neural Network. *Anal Chem*, pp. 2702-2709.

Zeng, L.U., Gu, Z., Fang, X., Fanwick. P.E., Chang, C., Smith, D.L. and McLaughlin, J.L. (1995). Two New Bioactive Triterpenoids from *Melia volkensii* (Meliaceae). *Tetrahedron*. **51**(9), 2477-2488.

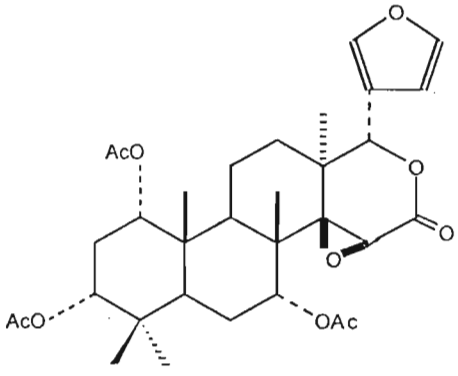
Zupan, J. and Gasteiger, J. (1991). *Anal. Chim. Acta*. **248**, 1-30.

Appendix I

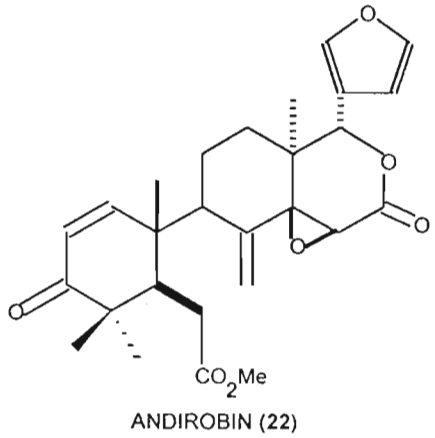
List of Structures



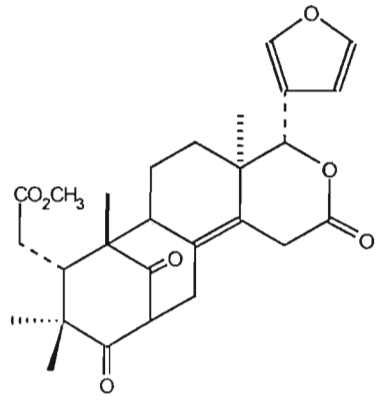
HAVANENSIN (20)



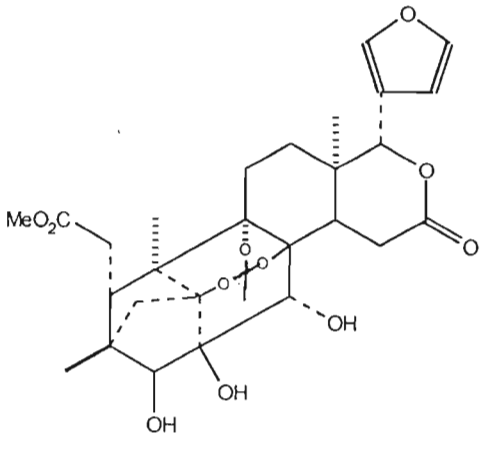
KHIVORIN (21)



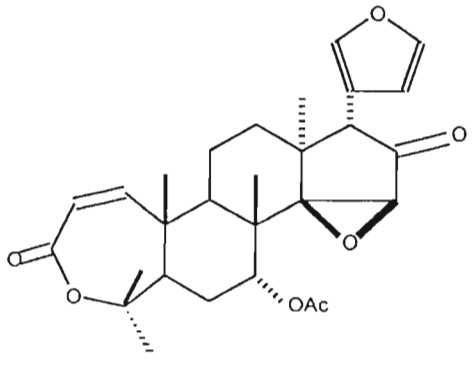
ANDIROBIN (22)



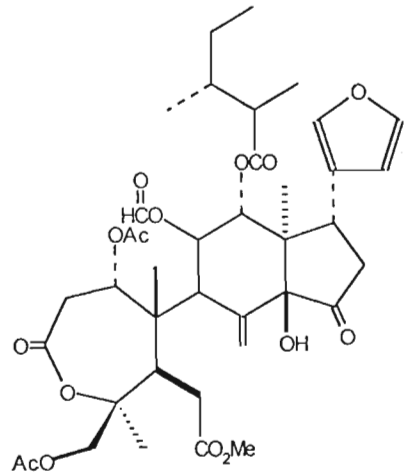
MEXICANOLIDE (23)



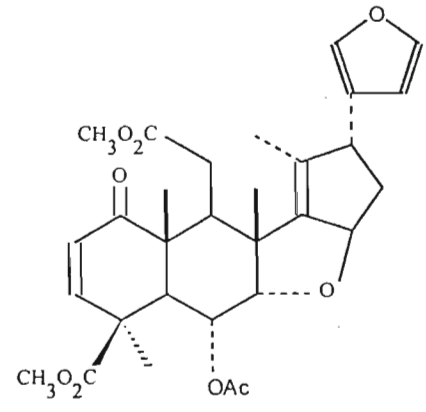
PHRAGMALIN (24)



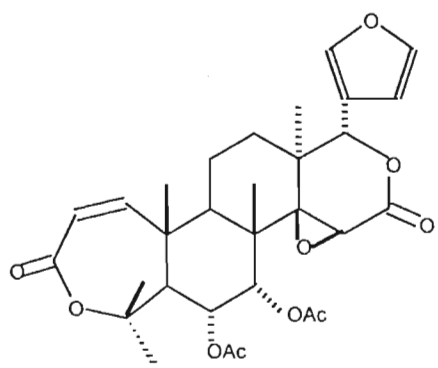
EVODULONE (25)



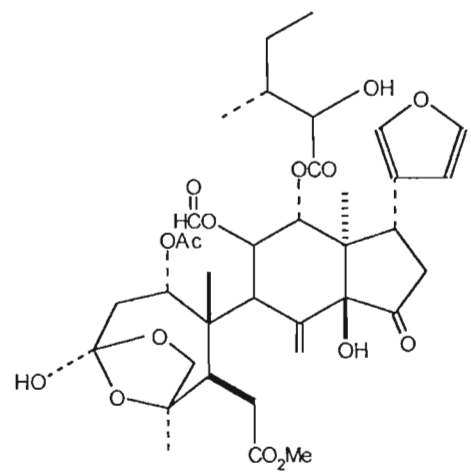
PRIEURIANIN (26)



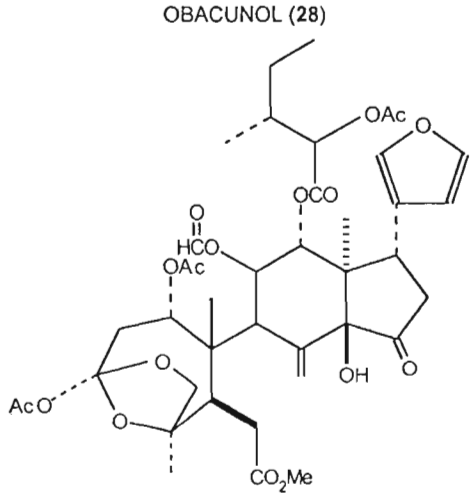
NIMBIN (27)



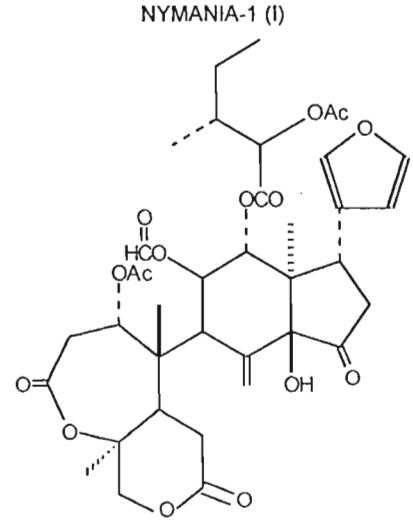
OBACUNOL (28)



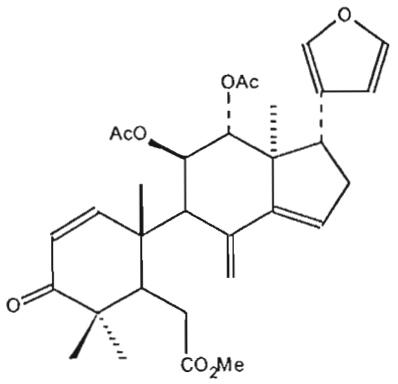
NYMANIA-1 (I)



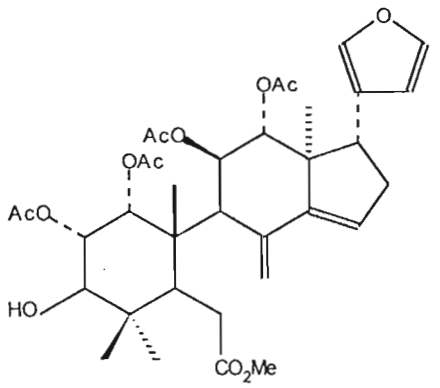
NYMANIA-1 ACETATE (II)



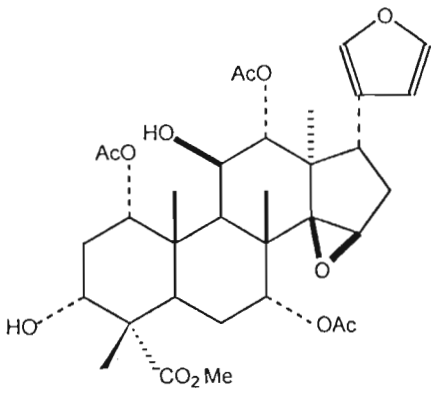
ROHITUKIN TYPE LIMONOID (III)



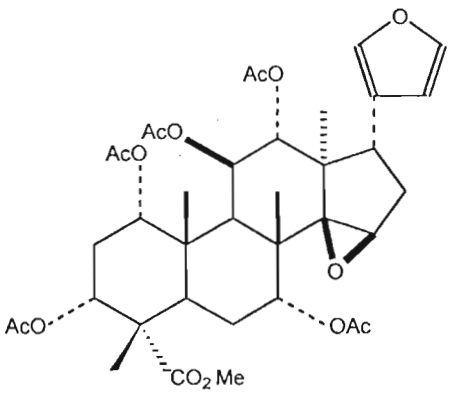
COMPOUND IV



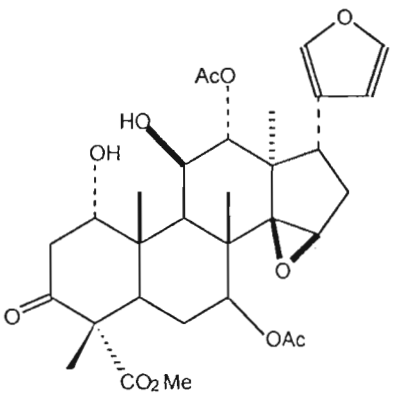
COMPOUND V



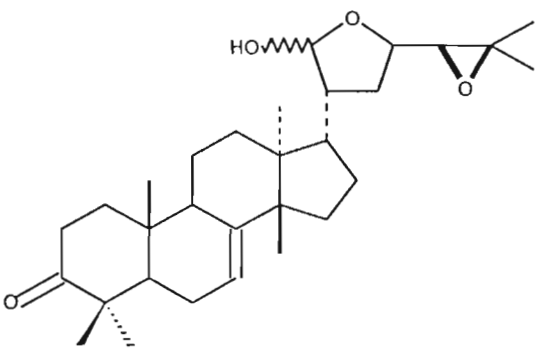
COMPOUND VI



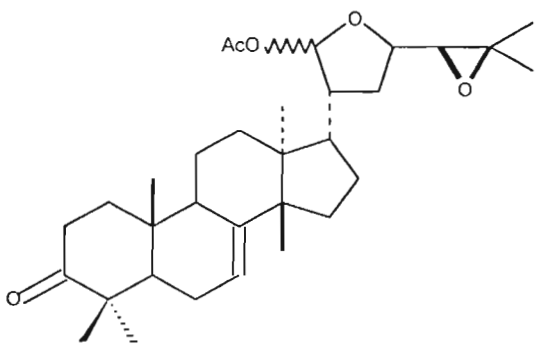
COMPOUND VII



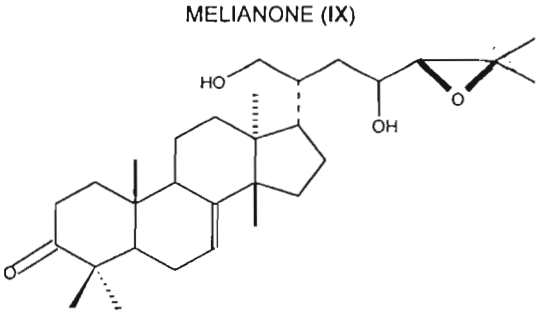
COMPOUND VIII



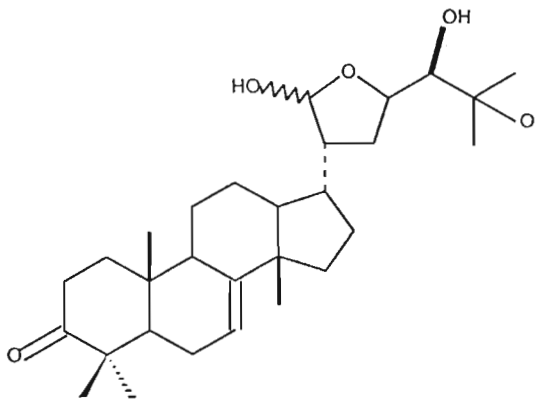
MELIANONE (IX)



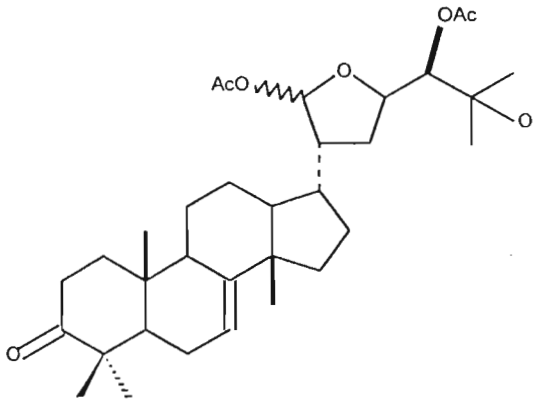
MELIANONE ACETATE (X)



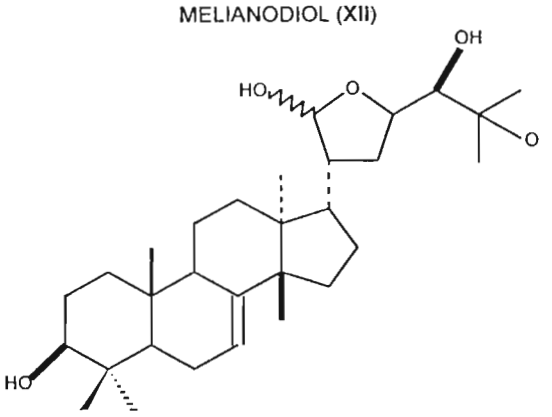
SAPELIN F (XI)



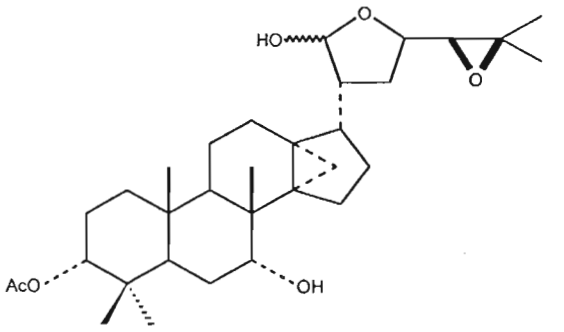
MELIANODIOL (XII)



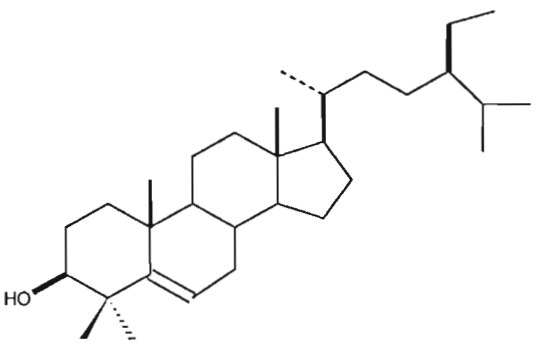
MELIANODIOL ACETATE (XIII)



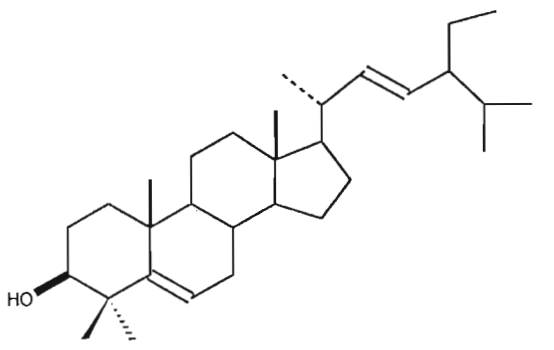
MELIANOTRIOL (XIV)



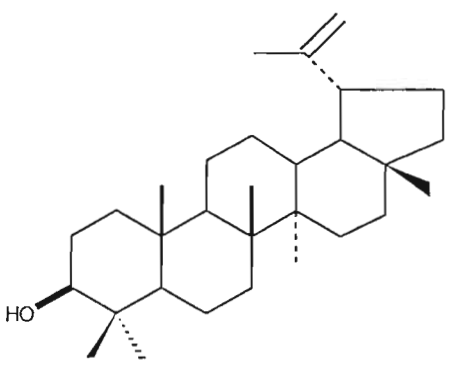
3α-ACETOXY-7α-DEACETOXY-GLABRETAL (XV)



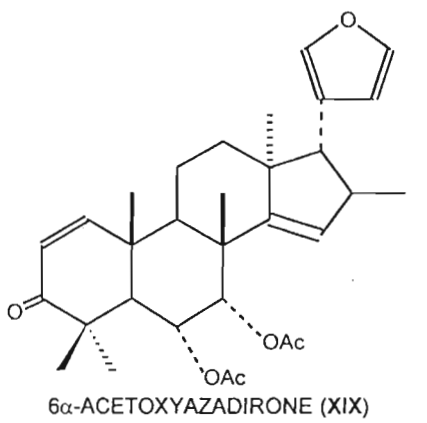
SITOSTEROL (XVI)



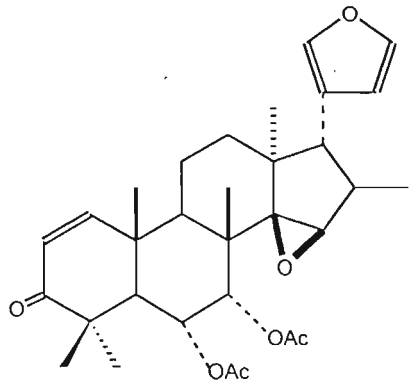
STIGMASTEROL (XVII)



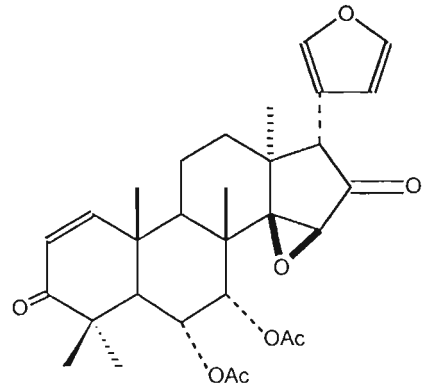
LUPEOL (XVIII)



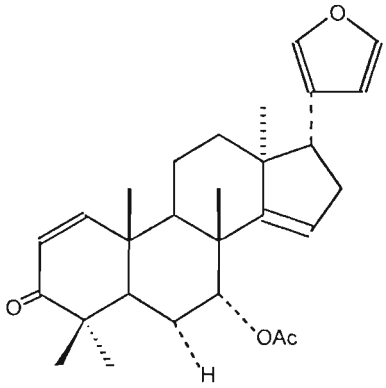
6α-ACETOXYAZADIRONE (XIX)



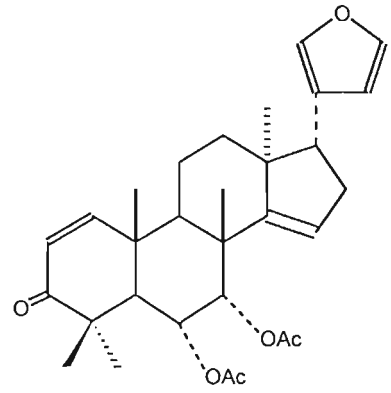
6α-ACETOXY-14β,15β-EPOXYAZADIRONE (**XX**)



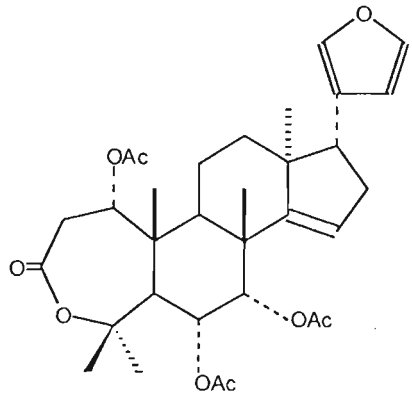
6α-ACETOXY-14β,15β-EPOXYAZADIRADIONE (**XXI**)



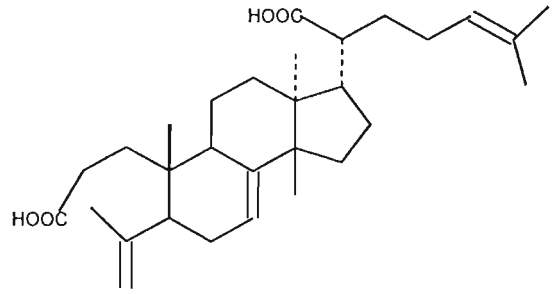
14β,15β-EPOXYAZADIRONE (**XXII**)



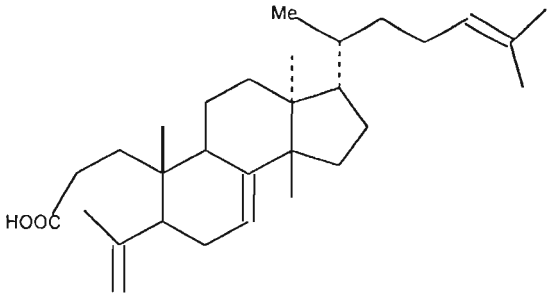
AZADIRONE (**XXIII**)



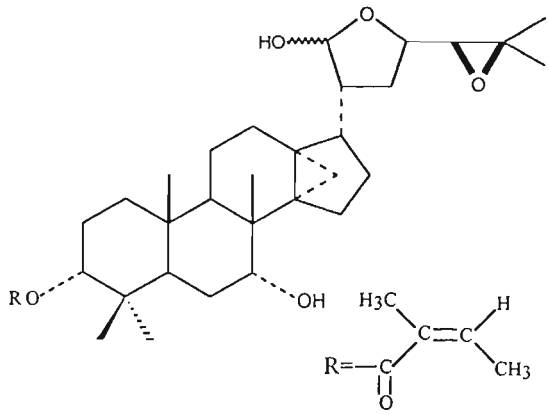
COMPOUND **XXIV**



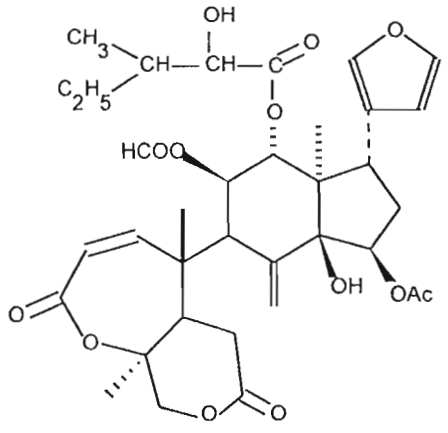
COMPOUND **XXV**



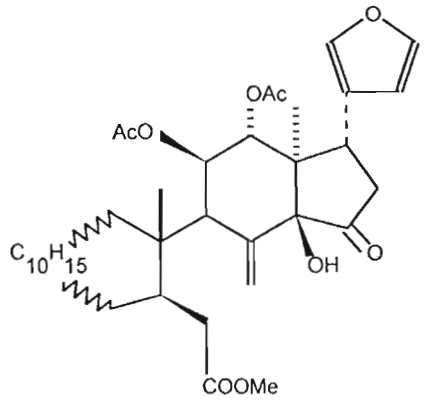
COMPOUND **XXVI**



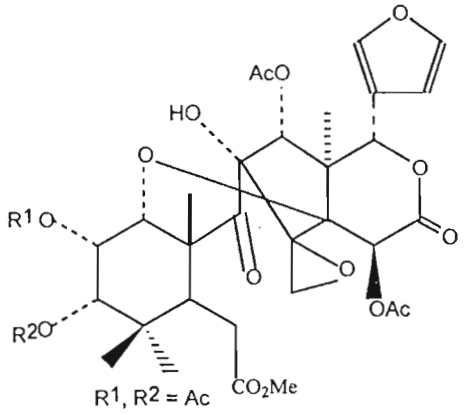
COMPOUND **XXVII**



COMPOUND XXVIII

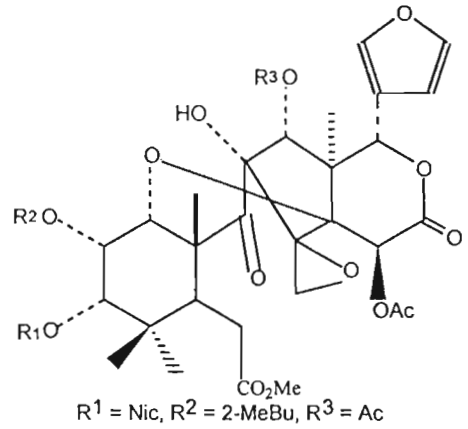


COMPOUND XXIX (structure could not be determined)



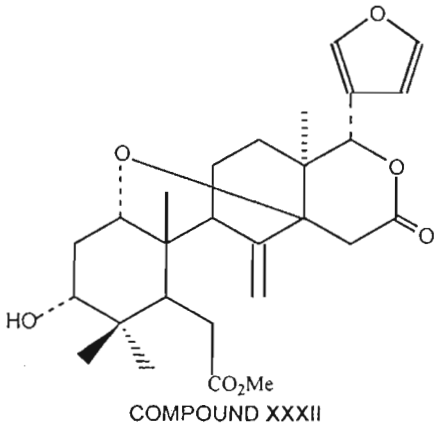
R^1, R^2 = Ac
or
R^1 = 2-MeBu, R^2 = Ac

COMPOUND XXX

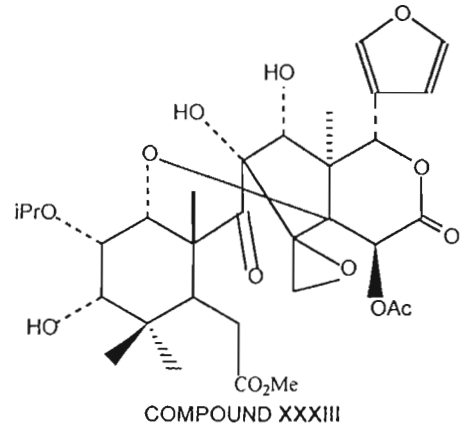


R^1 = Nic, R^2 = 2-MeBu, R^3 = Ac
or
R^1 = 2-MeBu, R^2 = Nic, R^3 = Ac

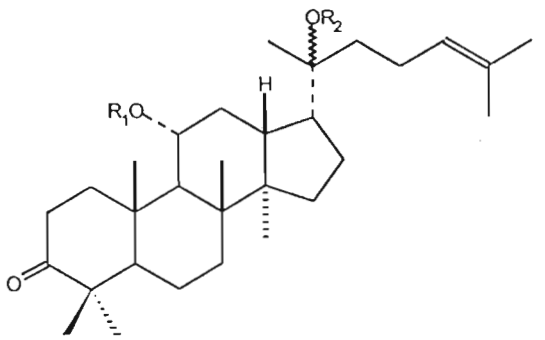
COMPOUND XXXI



COMPOUND XXXII

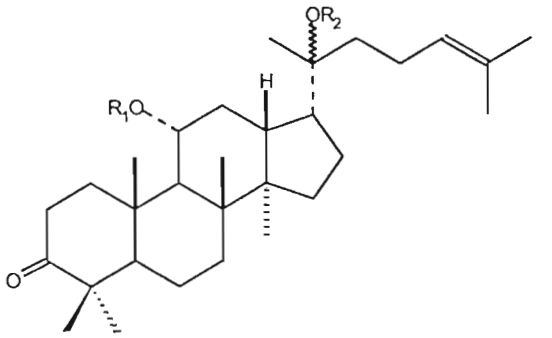


COMPOUND XXXIII



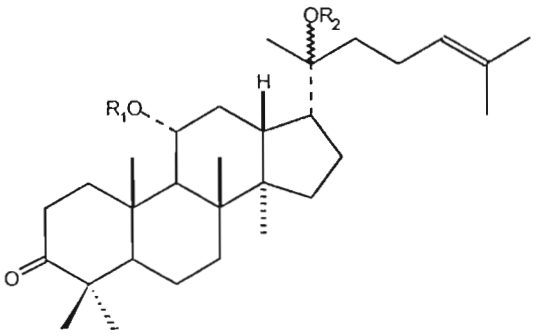
R_1	R_2
H	H

COMPOUND XXXIV



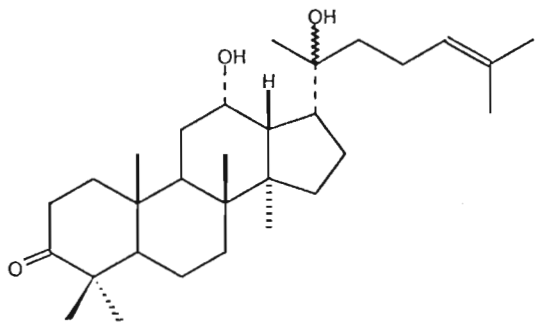
R_1	R_2
Ac	H

COMPOUND XXXV

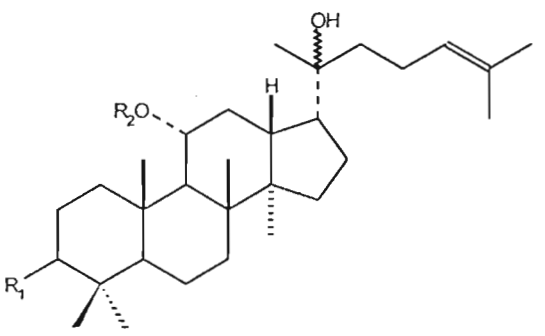


R_1	R_2
Ac	Ac

COMPOUND XXXVI

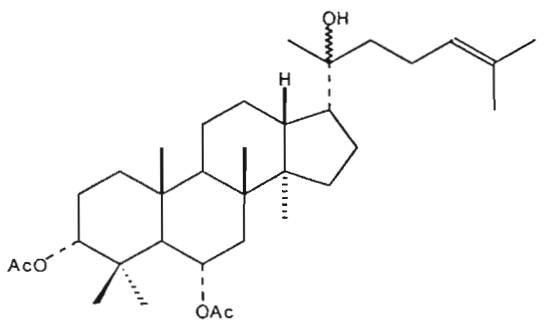


COMPOUND XXXVII

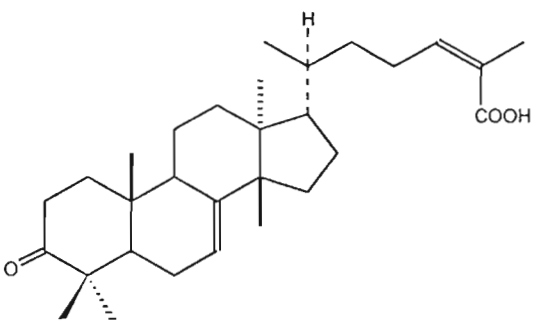


R_1	R_2
α-OH	H

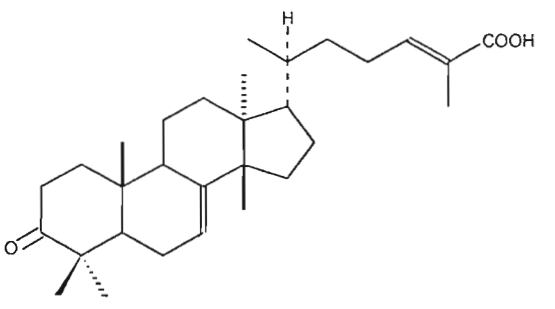
COMPOUND XXXVIII



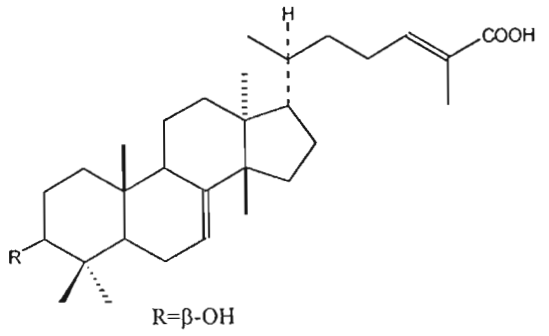
COMPOUND XXXIX



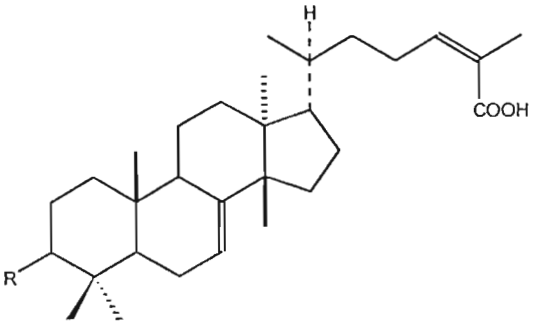
COMPOUND XXXX



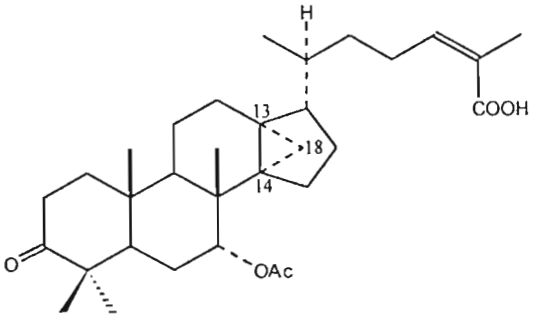
COMPOUND XXXXI



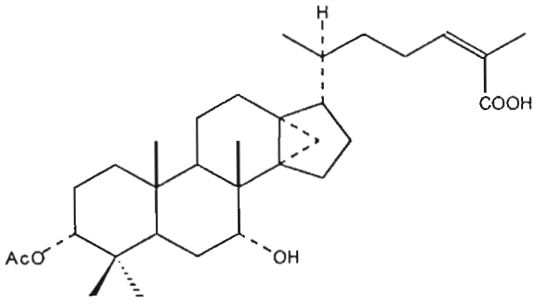
COMPOUND XXXXII



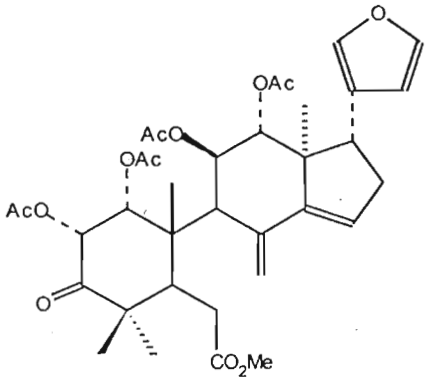
COMPOUND XXXXIII



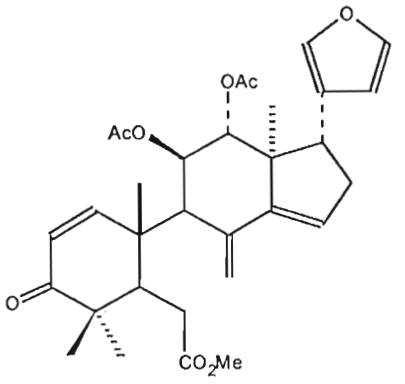
COMPOUND XXXXIV



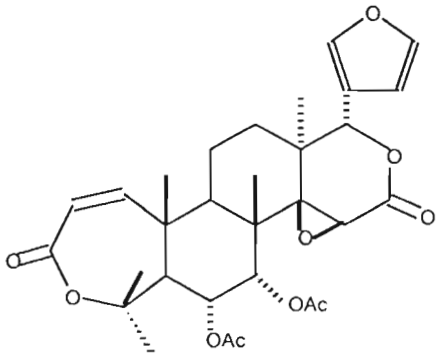
COMPOUND XXXXV



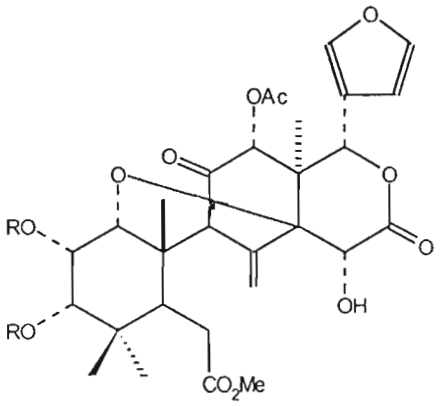
COMPOUND XXXXVI



COMPOUND XXXXVII

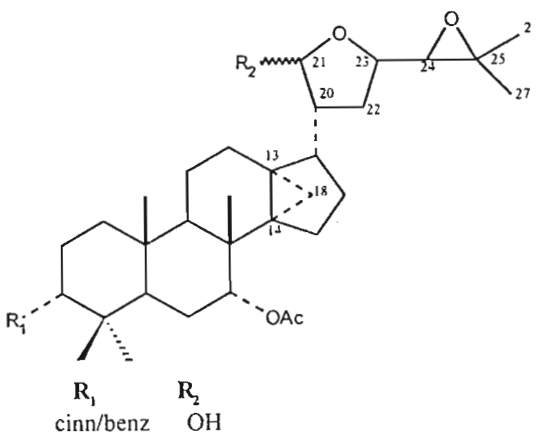


COMPOUND XXXXVIII



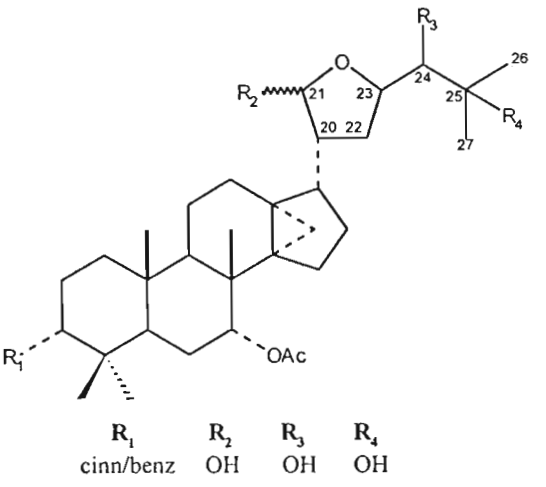
R = CINNAMATE OR NICOTINATE ESTER

COMPOUND XXXXIX



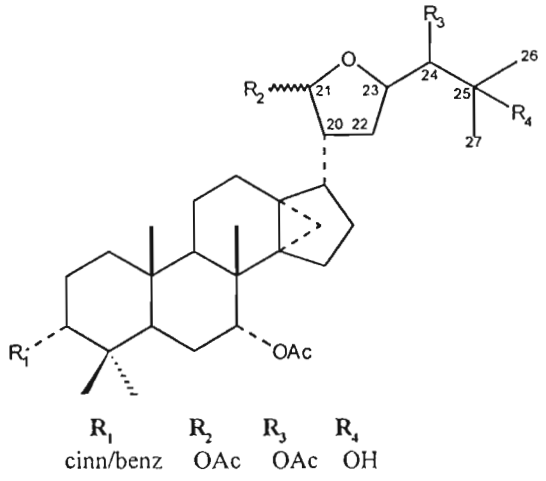
R_1	R_2
cinn/benz	OH

COMPOUND L



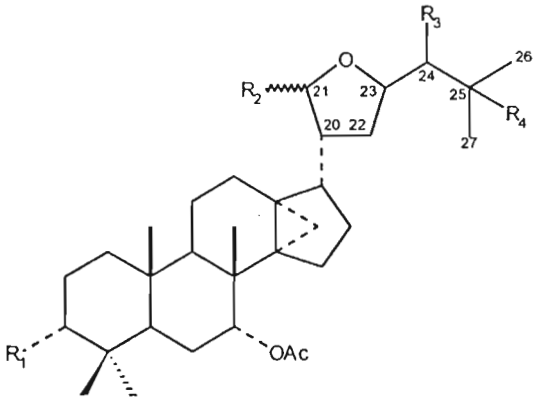
R_1	R_2	R_3	R_4
cinn/benz	OH	OH	OH

COMPOUND LI



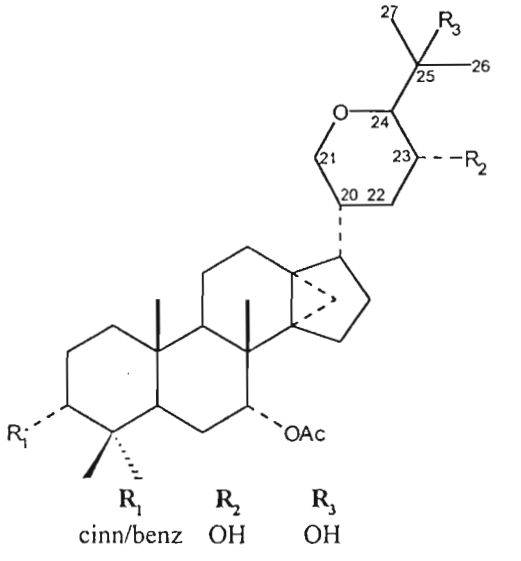
R_1	R_2	R_3	R_4
cinn/benz	OAc	OAc	OH

COMPOUND LII



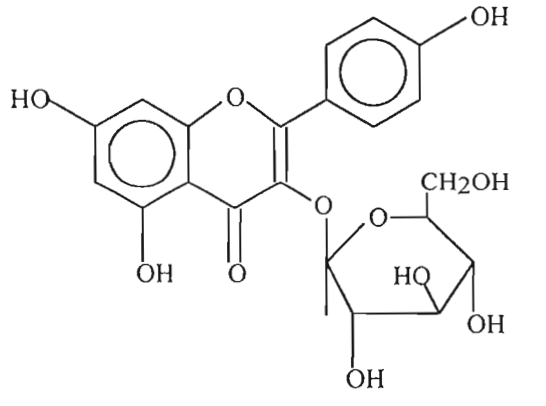
R_1	R_2	R_3	R_4
cinn/benz	OAc	OAc	OAc

COMPOUND LIII

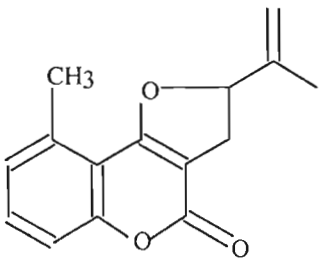


R_1	R_2	R_3
cinn/benz	OH	OH

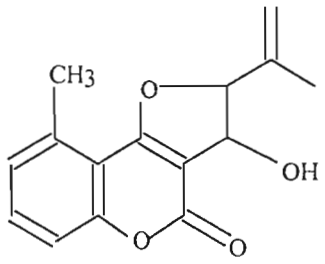
COMPOUND LIV



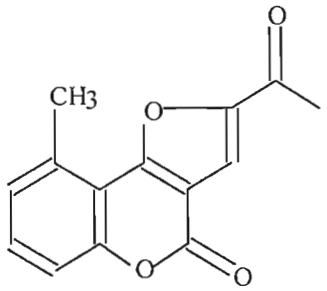
COMPOUND LV



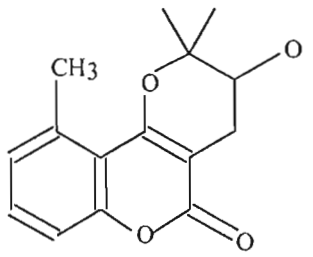
COMPOUND LVI



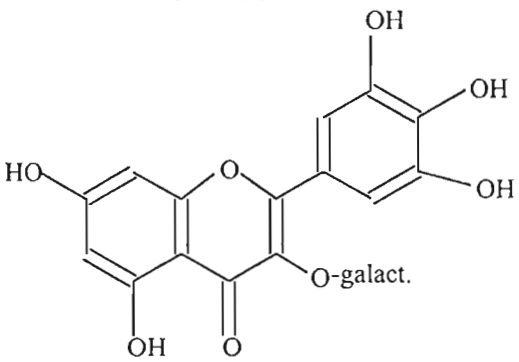
COMPOUND LVII



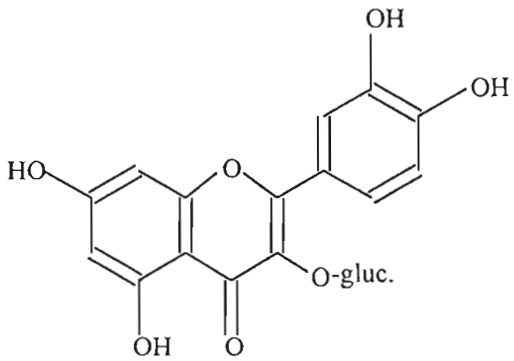
COMPOUND LVIII



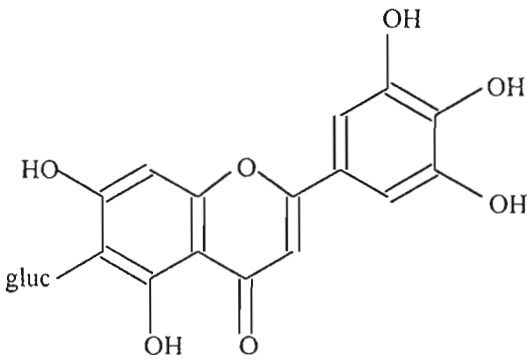
COMPOUND LIX



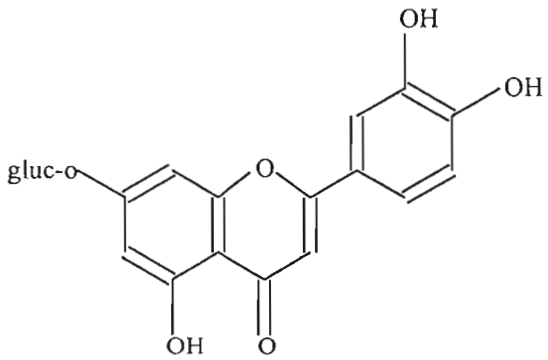
COMPOUND LX



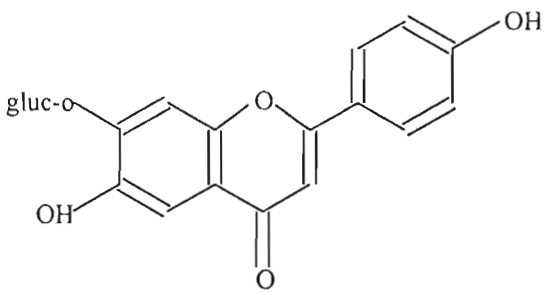
COMPOUND LXI



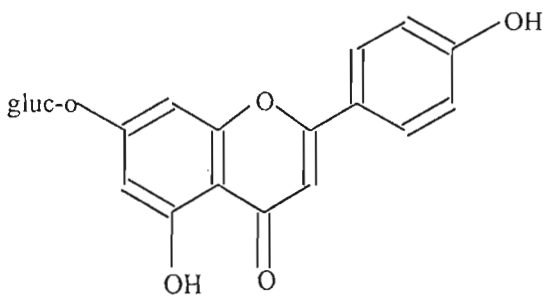
COMPOUND LXII



COMPOUND LXIII



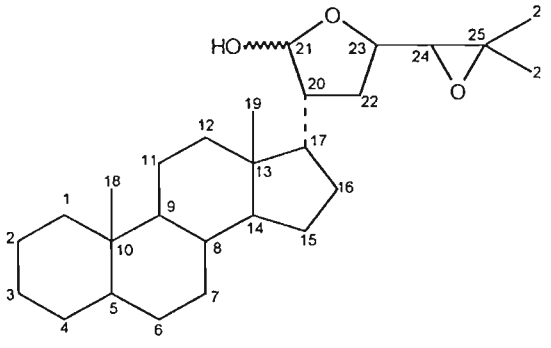
COMPOUND LXIV

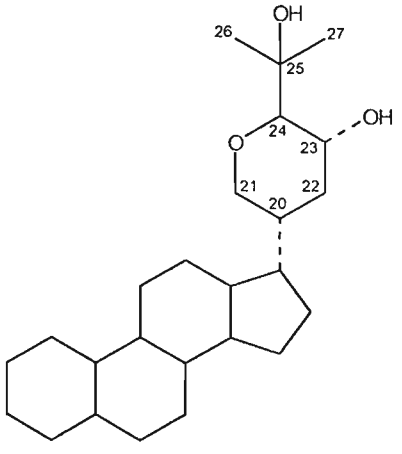


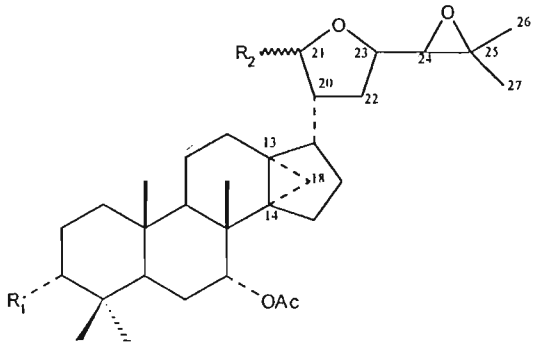
COMPOUND LXV

Appendix II

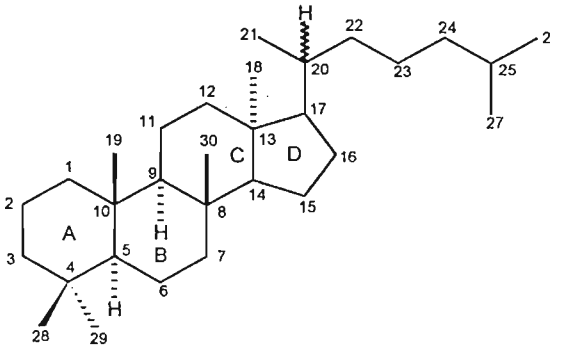
Numbering of Structures

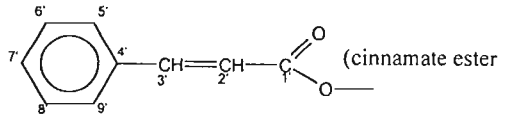




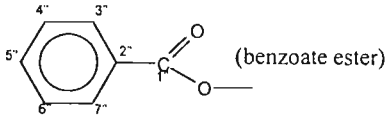


R_1	R_2
cinn/benz	OH
cinn/benz	OAc

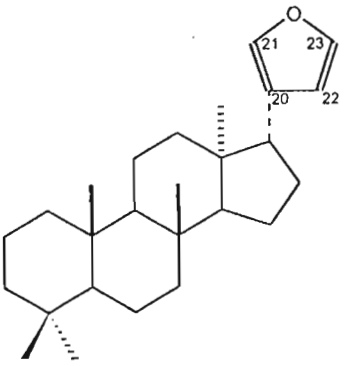


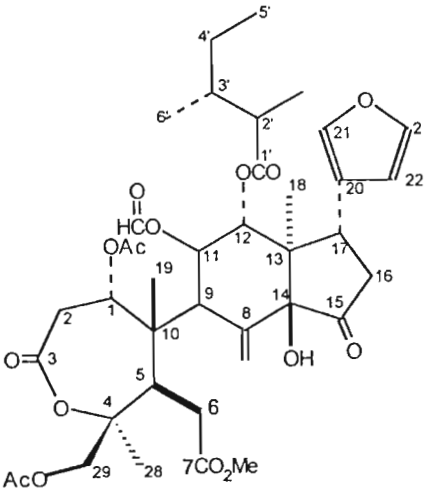


(cinnamate ester



(benzoate ester)


O
21
23
20
22


5'
4'
3'
6'
2'
1'
OCO
O
21
2
20
22
18
O
HCO
OAc
12
11
13
17
16
19
9
14
15
2
1
8
10
O
3
OH
O
5
O
4
6
AcO
29
28
7CO2Me

Appendix III

Experiments A, B and C

A. Classification of extra compounds

The peak positions of extra compounds described in Section 6.1 were entered into Microsoft Excel 97. The data entered was taken directly from the ^{13}C tables in the literature. No reliable (NMR spectrometer-sourced) amplitude data was present. The amplitudes were then either 100% (peak present) or 0% (no peak).

The data was binned using the identical procedures to those of the original training and test data sets. 5 ppm binning was used throughout. The neural network that determined the difference between limonoids, triterpenoids, flavonoids/coumarins and "others" was used for testing the performance of these unknown samples (taken incidentally using 400, 500 and 600 MHz, NMR spectrometers).

The neural network that had been already developed on the Meliaceae-based compounds was presented with the data. The neural network then determined the outputs, i.e. the estimation of the class of each new compound.

A.1 Results

Table III-1. Comparison of Actual versus Neural Network-Classified Classes

COMPOUND NAME	ACTUAL	CLASSIFIED
Bryon 1	t	t
Secobry4	t	t
Secoi 5	t	t
Meliav1a	t	t
Entandro	t	t
SapelinA	t	t
VilmorinineA	q	t
AilantinolA	q	t
AilantinolB	q	t
shinjudilactone	q	t
Ailanthone	q	t
olean1	t	t
olean2	t	t
yursen3	t	t
cincholic4	t	t
pyrocin5	t	t
pyro6	t	t
ref7	t	t
ref8	t	t
ref9	t	t
ref10	t	t
Xylocarpin(2)	l	l
RuageaninB(4)	l	l
RuageaninD(6)	l	l

Classification performance was 100% correct.

Where: q = quassinoids, t = triterpenoids, l = limonoids

A.2 Discussion & Conclusions

The results indicate that despite the compounds originating from another species, extracted by different processes, on different machines by different researchers, and without amplitude data, 100% correct identification was possible. It is important to note the relatively strong insensitivity of performance to amplitude data (or the lack of it). Also important to observe is that the quassinoids were correctly classified as triterpenoids rather than as "others". This shows that the neural network has developed a good general model of these classes allowing wide variability without misclassification.

B. Addition of random noise

The entire data set consisting of all the pure and impure compounds used in the original neural network developments was used for this experiment. As recommended in the reference report, noise was "randomly" added to each peak in increments of +/- 5% from 0-100%. This was achieved using the following equation:

$$P_{new} = P_{old} + (1 - 2R)\frac{Y}{100}P_{old} \qquad \text{(III-1)}$$

Where:

P_{new} = new peak amplitude

P_{old} = original peak amplitude

R = random number, evenly distributed over [0,1]

Y = percentage of noise added [0, 5, 10...100] = maximum noise added to signal

For example a peak of 100, with a random number of 0.1 and 50% noise would then change to 140.

As indicated in the earlier discussion, this process can at best only achieve a rough indication of performance under additive noise conditions and cannot be construed as statistically valid. The neural network that had been already developed without the added noise was presented with the noise added data. The neural network then determined the outputs, i.e. the estimation of the class of each compound. This was performed twice with different random "seed" values, i.e. different sets of random numbers.

B.1 Results

As the following plots from Microsoft Excel 97 indicate, there were between 0 and 11% misclassifications across the entire range of additive noise from 0-100% (Figure III.1 and III.2). The plots represented the two experiments in which the random seed was changed. The solid line in each plot indicates the trend of the results and was obtained by fitting a 2nd order polynomial to the results using Microsoft Excel.

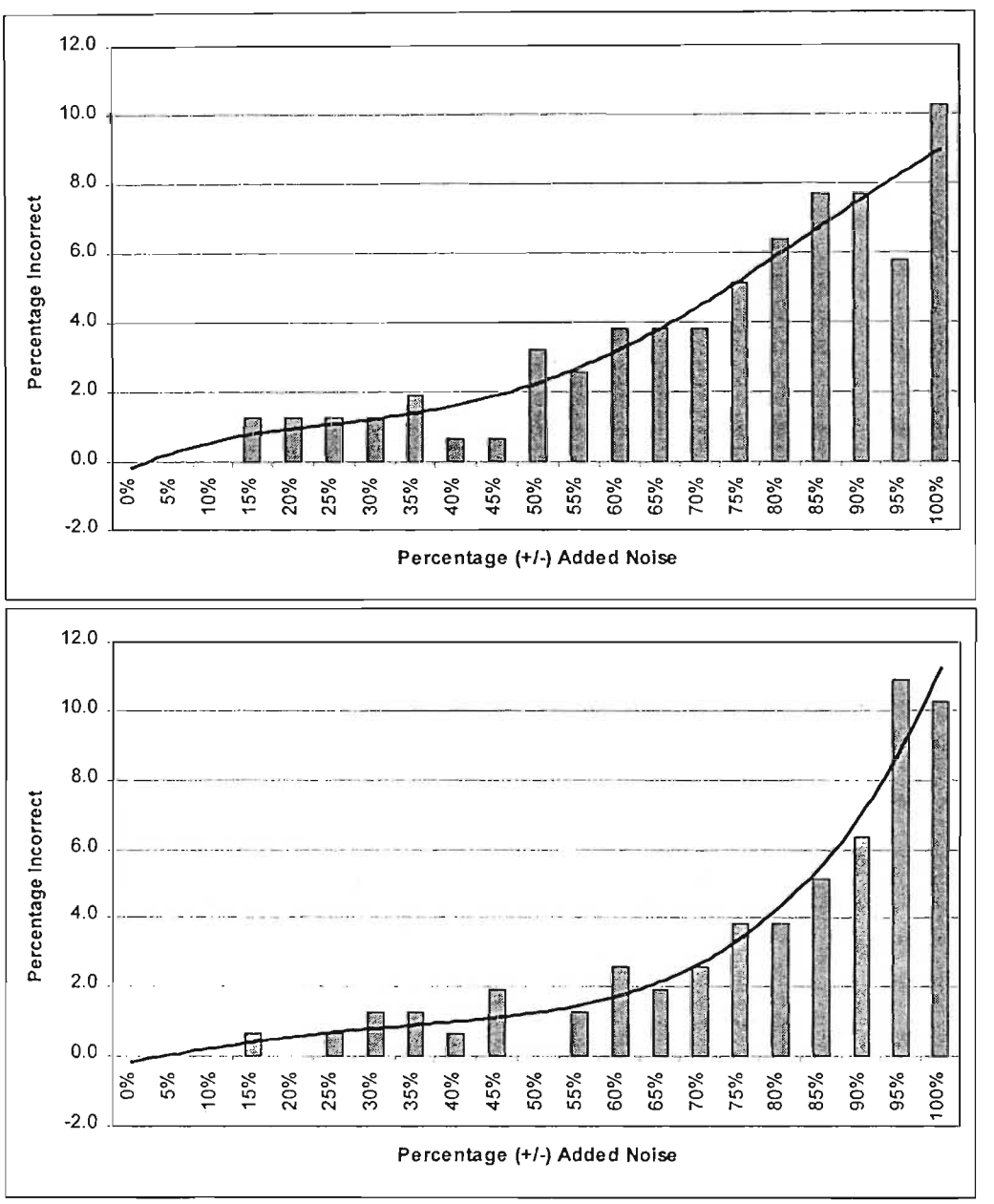


Figure III-1a&b Experimental results of the percentage of incorrectly classified compounds versus the percentage of additive noise.

B.2 Discussion & Conclusion

As the figures above indicate, the presence of noise appears to have an effect only at around 15-20% levels. The effect is the misclassification of 0.5 - 1.5% of compounds. Only at around 50-70% additive noise levels do the results become worse, but again not by much more than a couple of percent.

As discussed theoretically in Chapter 6, the effect of noise can have both positive and negative effects on the classification ability of the neural network as the results show. This can be particularly clearly seen at the 95% noise levels. The results can vary by several percent in absolute terms (and much more in relative terms) owing to this effect. This shows that to obtain statistically significant figures of merit for the impact of noise, exhaustive experimentation of prohibitive calculation duration would be required. The results merely give a broad qualitative range of possible influence of noise. The effect of randomness can be seen in the shape of the trend-lines that give considerably different results.

The strong result of this experiment is that the influence of amplitude on the decision-process, as already seen from **EXPERIMENT A**, is relatively small. The positional information of the bins is much more important.

C. Missing data analysis

A similar approach to **EXPERIMENT B** was taken for the "random" removal of each bin. A random number generator in Microsoft Excel 97 produced a number in the range [0,1] with a long-term even probability. This number was related to each bin using the following rule:

IF $R >= P$, THEN $V_x = V_x$
ELSE $V_x = 0$

(III-2)

Where
V_x = value in bin x
R = random number in [0,1]
P = probability of bin removal
x = bin number [0,1,2...44]

This process was relatively objective, but is however heavily skewed by the lack of sufficient experiments owing to impossibly long computational time. The neural network that had been already developed on the Meliaceae-based compounds was presented with the data three times for each probability of missing bin. The neural network then determined the outputs, i.e. the estimation of the class of each new compound.

C.1 Results

The results, as expected, varied widely as the random numbers changed. Figure III-2 shows the average results of the three experiments per missing bin probability.

When the probability was small, i.e. 0.1 (or 10%), the effect on classification was just over 6% (average). This rose to around 26% for a 0.5 probability of a bin missing. The trend line shows an approximately linear slope to for this effect.

The results tend to indicate that the probability of bin missing is proportional to twice the misclassification rate.

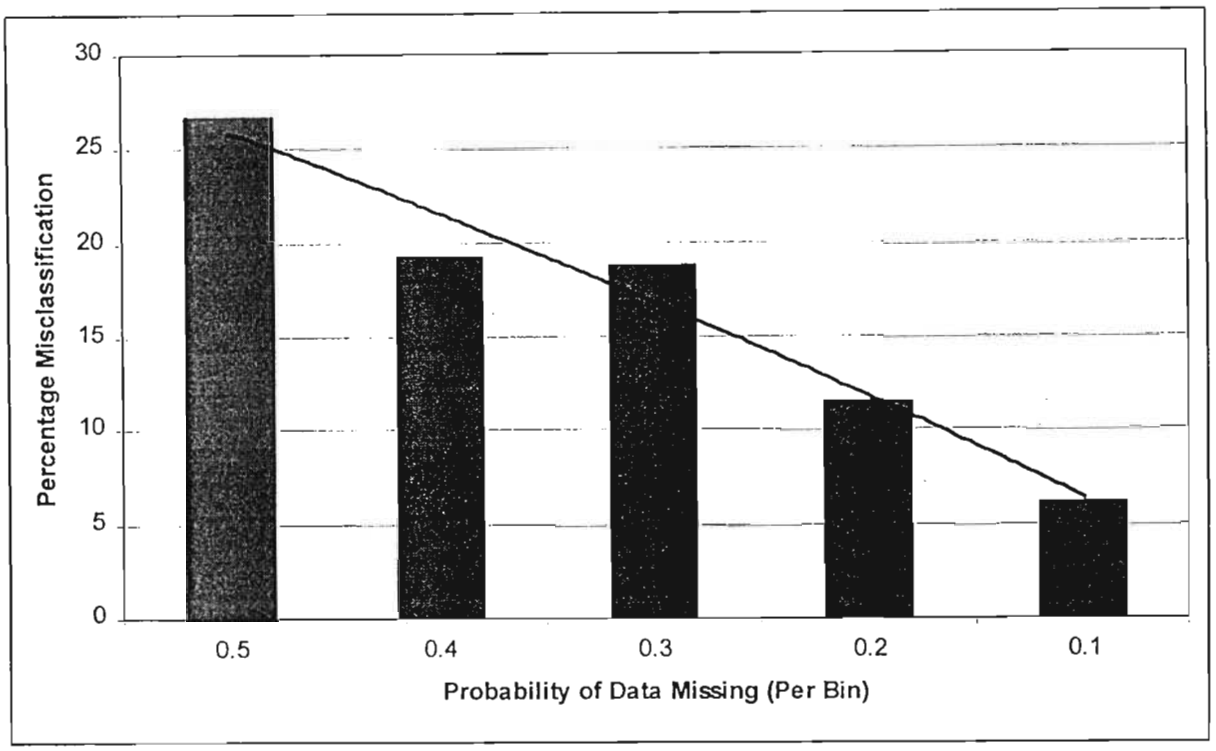


Figure III-2 Experimental results of the percentage of incorrectly classified compounds versus the probability of each bin to be missing.

Table III-2. Results of Experiment C

Compound	Target	Prob. = 0.5			prob. = 0.4			prob. = 0.3			prob. = 0.2			prob. = 0.1		
		Experiment			Experiment			Experiment			Experiment			Experiment		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
		result			result			result			result			result		
1	I	f/c	I	t	t	f/c	I	I	I	I	I	I	I	t	I	I
2	I	t	t	t	I	I	I	I	I	I	I	t	I	I	I	I
3	I	f/c	f/c	I	I	I	I	f/c	f/c	I	I	I	I	I	I	I
4	I	t	I	f/c	I	I	t	t	I	I	I	I	I	I	I	I
5	I	I	f/c	I	I	I	I	I	I	I	t	I	t	I	I	I
6	I	I	I	t	I	I	t	t	I	I	I	I	I	I	I	I
7	I	t	t	f/c	I	I	t	t	t	I	I	I	I	I	I	t
8	I	t	I	f/c	f/c	f/c	t	f/c	f/c	t	t	I	I	I	I	I
9	I	t	I	t	t	f/c	I	I	t	I	t	I	I	I	I	I
10	I	f/c	I	I	I	f/c	I	I	t	I	I	t	I	I	I	I
11	I	I	I	I	f/c	I	t	I	I	I	I	t	I	I	I	I
12	I	I	t	I	f/c	I	I	t	I	I	I	I	I	I	I	I

13	l	l	f/c	t	f/c	l	t	t	l	l	l	l	t	l	l	l
14	l	l	l	f/c	t	l	t	t	f/c	l	l	t	l	l	l	l
15	l	l	f/c	f/c	f/c	t	l	t	f/c	l	t	t	f/c	f/c	l	l
16	l	f/c	l	l	l	l	l	f/c	t	t	l	l	l	l	l	l
17	l	t	l	l	t	l	l	f/c	f/c	l	l	l	f/c	l	l	l
18	l	t	t	f/c	f/c	t	l	l	f/c	l	t	l	l	t	l	l
19	l	l	l	l	t	t	t	l	t	l	t	t	l	l	l	l
20	l	f/c	l	l	l	t	t	l	l	t	t	l	t	t	t	l
21	l	f/c	f/c	f/c	f/c	f/c	f/c	l	l	l	l	l	t	t	l	l
22	l	l	l	f/c	l	l	l	l	l	l	t	l	l	l	l	l
23	t	t	t	f/c	f/c	t	f/c	t	t	t	t	t	t	t	t	t
24	t	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t
25	t	t	t	f/c	f/c	t	t	t	t	f/c	t	f/c	t	t	t	t
26	t	l	t	l	l	t	t	t	t	f/c	t	l	t	l	t	l
27	t	t	f/c	t	t	f/c	f/c	f/c	t	t	t	t	t	t	t	t
28	t	f/c	t	f/c	t	t	f/c	f/c	t	t	t	t	f/c	t	t	t
29	t	t	t	f/c	t	t	f/c	t	t	f/c	t	t	f/c	t	t	t
30	t	f/c	t	t	t	f/c	f/c	t	t	t	t	t	t	t	t	t
31	t	f/c	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t
32	t	t	t	t	t	f/c	f/c	t	f/c	t	t	t	t	t	t	t
33	t	t	f/c	f/c	t	t	f/c	t	t	f/c	t	t	t	t	t	f/c
34	t	t	f/c	t	f/c	t	t	t	t	f/c	t	f/c	t	t	t	f/c
35	t	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	f/c
36	t	t	t	t	t	t	t	t	t	t	f/c	t	t	t	t	t
37	t	f/c	t	t	t	t	t	f/c	t	t	t	t	t	t	t	t
38	t	t	t	t	f/c	t	f/c	f/c	l	l	t	t	t	t	t	t
39	t	t	t	l	t	t	t	l	t	l	t	l	t	t	t	t
40	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t
41	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t
42	t	f/c	f/c	f/c	f/c	t	t	t	t	t	t	t	t	t	t	t
43	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
44	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t
45	t	t	t	t	f/c	t	t	t	t	f/c	t	t	f/c	t	t	t
46	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
47	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
48	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
49	t	f/c	f/c	t	f/c	t	t	t	t	t	t	t	t	t	t	t
50	t	l	f/c	t	f/c	t	t	l	l	t	t	t	t	t	t	t
51	t	f/c	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t
52	t	t	t	t	t	f/c	t	l	t	t	t	t	t	l	t	t
53	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
54	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
55	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
56	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
57	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
58	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
59	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
60	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
61	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
62	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
63	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
64	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c

65	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
66	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
67	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
68	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
69	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
70	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
71	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
72	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
73	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
74	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
75	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
76	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
77	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
78	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
79	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
80	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
81	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
82	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
83	l	t	f/c	l	f/c	l	l	l	f/c	f/c	l	l	l	l	l	l
84	l	t	t	l	l	t	t	t	l	l	l	l	l	l	t	f/c
85	l	l	l	l	t	l	l	l	l	l	t	l	l	l	l	t
86	l	t	l	l	l	t	t	f/c	l	l	l	l	l	l	l	l
87	l	l	l	f/c	l	f/c	l	f/c	f/c	l	l	l	l	l	l	l
88	l	t	t	t	t	l	l	t	t	l	l	t	l	t	l	l
89	l	f/c	l	l	f/c	l	l	t	l	l	t	t	l	l	l	l
90	l	f/c	l	l	l	l	l	l	l	f/c	t	l	l	l	l	l
91	l	l	f/c	l	f/c	l	l	l	l	l	t	f/c	t	t	l	l
92	l	t	t	t	l	l	l	t	l	l	l	l	l	t	l	l
93	l	l	l	l	l	l	f/c	l	l	l	l	t	l	l	l	l
94	l	l	t	t	l	l	t	l	f/c	t	t	l	l	t	l	l
95	l	l	t	f/c	l	t	l	t	t	t	l	l	f/c	l	l	t
96	l	l	t	l	f/c	l	f/c	t	t	t	l	l	l	t	l	t
97	l	t	l	l	f/c	f/c	l	f/c	t	l	f/c	l	l	l	l	l
98	l	f/c	l	l	t	l	l	l	l	l	l	t	t	t	l	l
99	l	f/c	t	f/c	f/c	l	l	t	f/c	l	l	l	l	l	l	l
100	l	f/c	f/c	t	l	f/c	l	t	l	f/c	l	t	l	l	l	t
101	l	t	l	f/c	t	l	f/c	f/c	l	t	t	l	l	l	l	t
102	l	t	t	t	l	l	t	l	l	t	t	l	l	l	l	l
103	l	l	l	f/c	l	l	t	l	l	l	l	l	l	l	l	l
104	l	f/c	f/c	t	f/c	t	l	l	l	l	t	l	t	l	l	l
105	t	t	t	f/c	t	f/c	t	f/c	t	t	t	t	t	t	t	t
106	t	t	t	t	t	t	t	f/c	t	t	t	t	t	t	t	t
107	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
108	t	l	t	f/c	t	l	t	l	t	t	t	t	t	t	t	t
109	t	f/c	t	f/c	f/c	t	t	t	t	f/c	t	t	t	t	t	f/c
110	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	f/c	t
111	t	t	t	t	t	t	t	t	l	t	t	t	t	t	t	t
112	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
113	t	t	t	t	t	t	t	t	t	f/c	t	t	t	t	t	t
114	t	f/c	t	t	f/c	t	t	t	t	t	t	f/c	f/c	t	t	t
115	t	f/c	t	f/c	f/c	t	t	t	t	f/c	t	t	t	t	t	t
116	t	t	f/c	t	t	t	t	t	t	t	f/c	t	t	t	t	t

117	t	t	t	t	f/c	t	t	f/c	f/c	t	t	t	t	t	t	t
118	t	t	f/c	t	t	l	l	t	l	f/c	t	t	t	t	t	t
119	t	t	f/c	t	t	t	t	t	t	t	f/c	f/c	t	t	t	t
120	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t	t
121	t	t	t	t	t	f/c	t	t	f/c	t	t	t	t	t	t	t
122	t	t	f/c	f/c	t	t	t	f/c	t	f/c	t	t	t	t	t	t
123	t	t	t	t	t	t	t	f/c	t	t	t	t	t	t	t	t
124	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t	t
125	t	t	f/c	t	f/c	t	t	t	f/c	t	t	t	t	t	t	t
126	t	f/c	t	t	t	f/c	t	t	t	t	t	t	t	f/c	t	t
127	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t	t
128	t	t	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t
129	t	l	t	t	t	t	t	t	t	l	t	t	t	t	t	t
130	t	t	t	f/c	t	t	t	t	t	t	t	t	t	t	t	t
131	t	f/c	f/c	f/c	t	t	t	t	t	t	f/c	t	l	t	t	t
132	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
133	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
134	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
135	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
136	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
137	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
138	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
139	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
140	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
141	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
142	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
143	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
144	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
145	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
146	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
147	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
148	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
149	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
150	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
151	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
152	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
153	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
154	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
155	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c
156	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c	f/c

C.2 Discussion & Conclusion

It is firstly important to notice that a probability of bin removal of 10% results in a loss of classification accuracy of 5 % (approximately) *not* 10% as may be surmised. This means that the neural network is robust in the presence of loss of information, yielding a better return than the rate of loss would presume.

Secondly, in comparing the results of **EXPERIMENT B** and **EXPERIMENT C**, the results show much greater sensitivity of the neural networks to positional loss of information than amplitude loss of information. This is exactly what would be expected, bearing in mind the performance of **EXPERIMENT A** on new compounds without explicit amplitude data.

Thirdly, the results are completely expected since the removal of bins is akin to presenting vastly different vectors to the neural network which it has never before experienced, and widely outside the area bounded by the training data set (as Figure 7 shows). In these synthetic circumstances, the performance of the neural networks is bound to be degraded owing to extrapolation rather than interpolation.

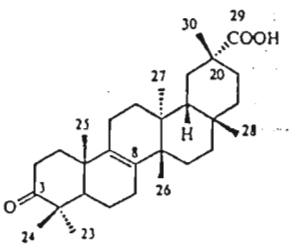
Finally, this experiment must be viewed as exceptionally harsh in that the probability of removal applies to *each* bin in the spectrum. This is highly unlikely in practice where the loss of data rarely affects all bins equally. In fact the loss of information is relatively unlikely at all since NMR thresholding coupled with the nuclear Overhauser effect serves merely to alter or emphasise individual peaks rather than to eliminate them completely. What is more likely is the presence of undesired additive noise peaks from the plant matrix.

Appendix IV

Bryononic Acid [10]

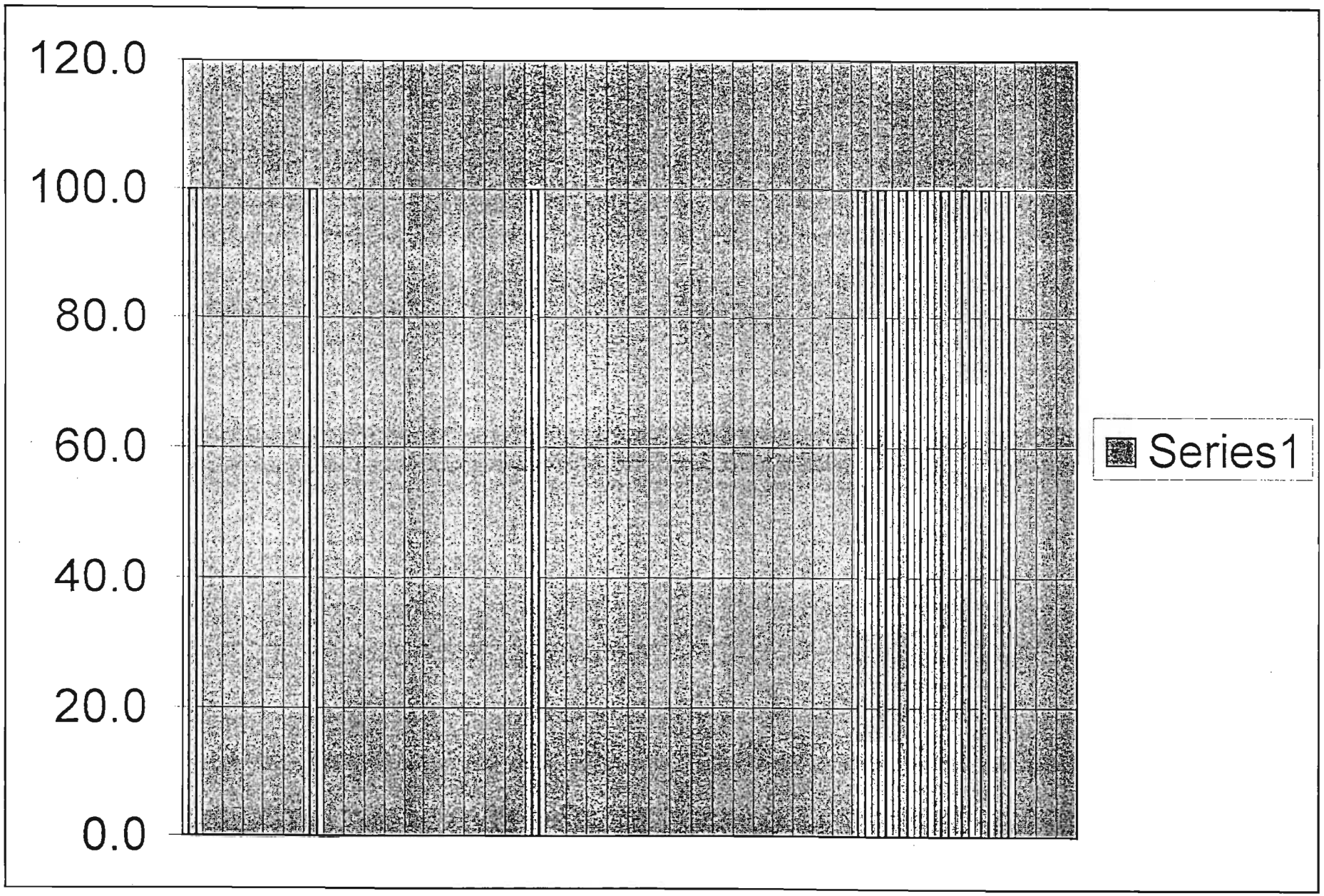
^{13}C NMR Data: (ppm)

218.2	185.5	134.9	132.7	51.1	47.1
44.4	42.2	40.4	37.4	37.0	36.8
35.4	34.4	34.2	32.7	31.2	30.8
30.4	29.9	29.5	27.7	26.8	25.2
21.6	21.1	20.6	20.5	19.4	18.1



Bryononic acid (1)

[10] Kosela, S., Yulizar, Y., Chairul, C., Tori, M., & Asakawa, Y., "Secomultiflorane-type triterpenoid acids from stem bark of Sandoricum koetjape", Phytochem., Elsevier, Great Britain, Vol. 38, no. 3, pp691-694, 1995.

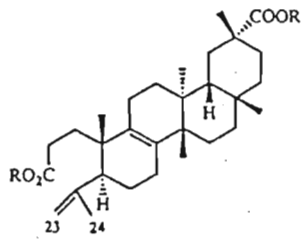


5 ppm Binned Spectrum of Bryononic Acid

Secobryononic Acid [10]

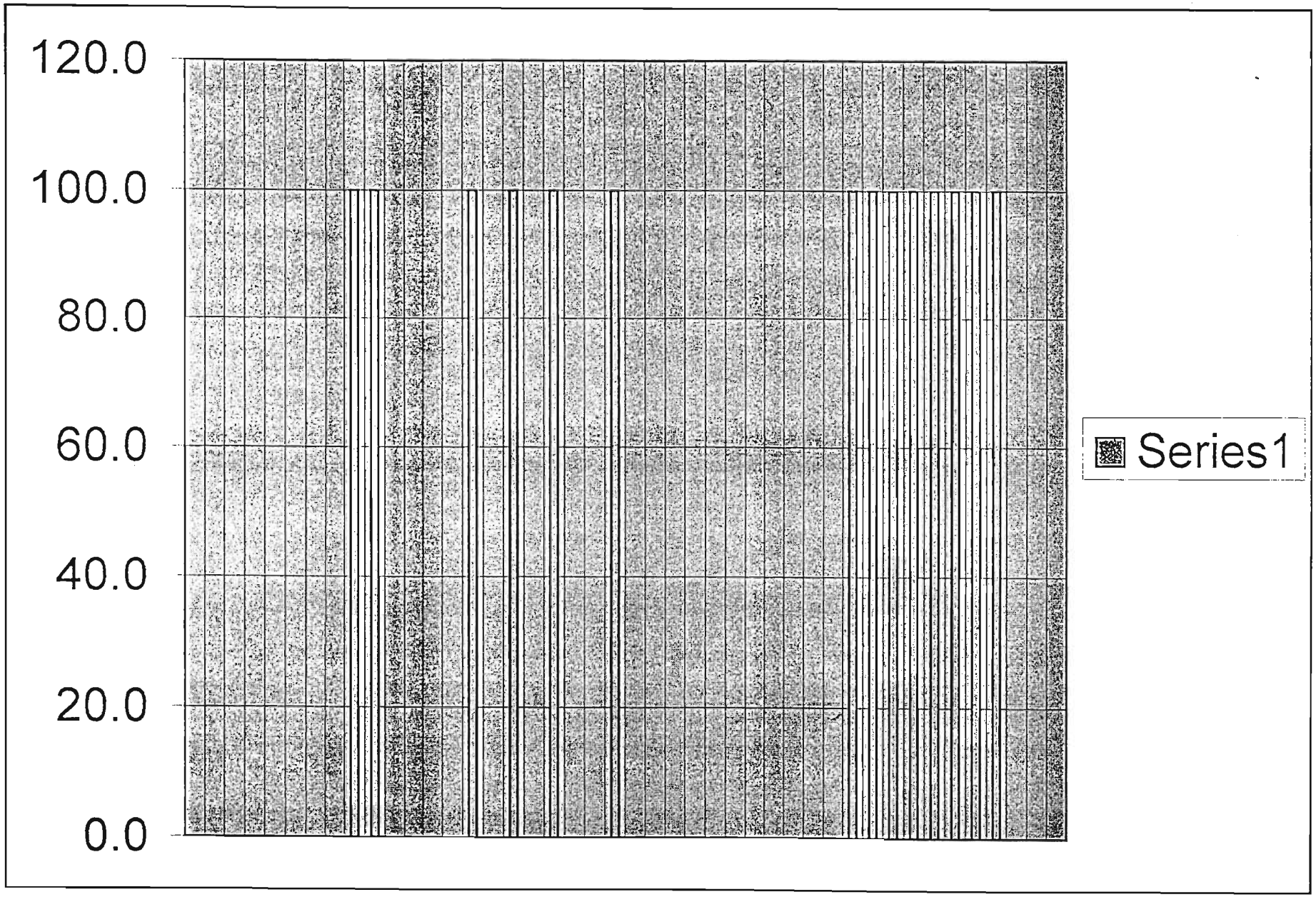
^{13}C NMR Data: (ppm)

179.2	174.7	147.4	138.9	129.5	113.8	51.5
51.4	47.8	46.4	44.6	41.2	40.4	37.1
36.9	34.4	32.7	31.6	31.3	30.9	30.7
30.3	30.0	29.7	26.5	25.1	24.8	23.3
23.1	21.6	20.9	18.1			



Secobryononic acid (2) R=H

[10] Kosela, S., Yulizar, Y., Chairul, C., Tori, M., & Asakawa, Y., "Secomultiflorane-type triterpenoid acids from stem bark of Sandoricum koetjape", Phytochem., Elsevier, Great Britain, Vol. 38, no. 3, pp691-694, 1995.

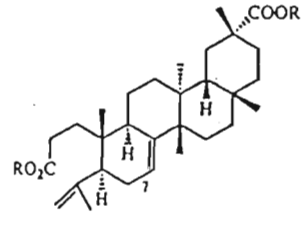


5 ppm Binned Spectrum of Secobryononic Acid

Secoisobryononic Acid [10]

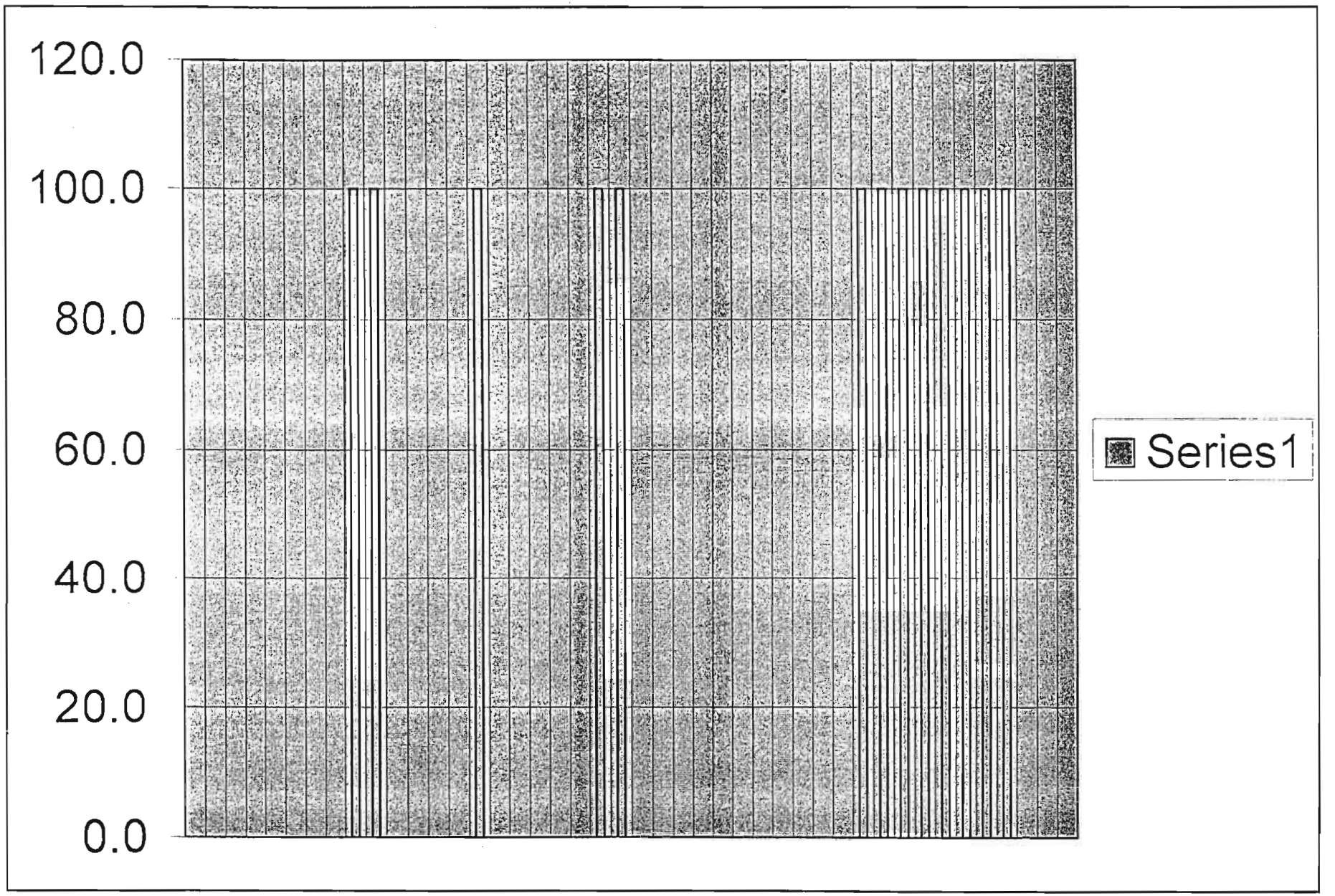
^{13}C NMR Data: (ppm)

179.4	174.8	147.5	145.7	117.2	113.8	51.4
51.4	49.0	47.3	42.5	40.4	40.3	37.1
36.9	36.0	35.7	33.2	32.9	31.7	31.4
31.2	30.7	30.3	29.6	29.2	28.1	24.8
23.9	22.5	17.5	15.9			



Secoisobryononic acid (3) R=H

[10] Kosela, S., Yulizar, Y., Chairul, C., Tori, M., & Asakawa, Y., "Secomultiflorane-type triterpenoid acids from stem bark of Sandoricum koetjape", Phytochem., Elsevier, Great Britain, Vol. 38, no. 3, pp691-694, 1995.

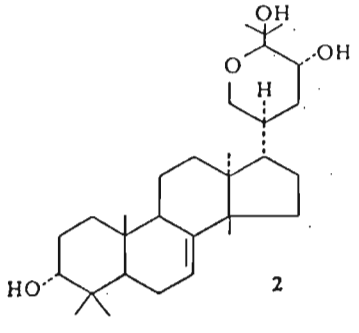


5 ppm Binned Spectrum of Secoisobryononic Acid

Sapelin A [8]

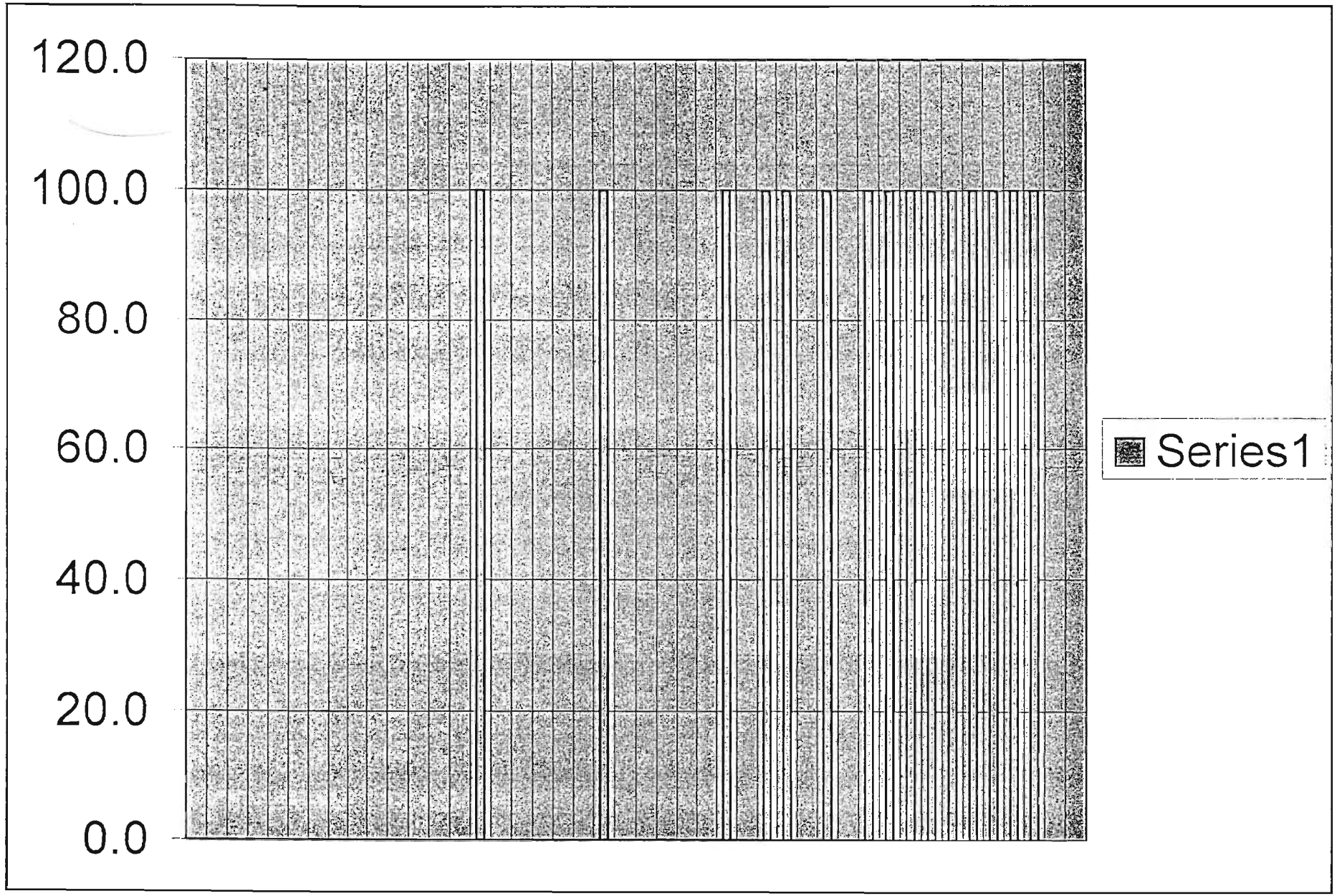
^{13}C NMR Data: (ppm)

145.8	118.2	86.5	76.3	74.1	70.2	64.7
51.4	48.7	48.7	44.8	44.6	43.4	37.4
36.5	34.8	33.9	33.1	31.3	28.4	27.8
27.4	27.3	25.5	24.0	24.0	22.3	21.9
17.0	13.0					



2

[8] Okorie, D.A. & Taylor, D.A.H., "Triterpenes from the seed of *Entandrophragma* species", *Phytochem.*, Pergamon Press, Great Britain, Vol. 16, pp2029-2030, 1977.

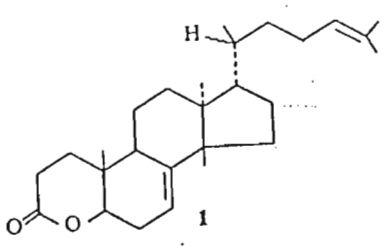


5 ppm Binned Spectrum of Sapelin A

Entandrolide [8]

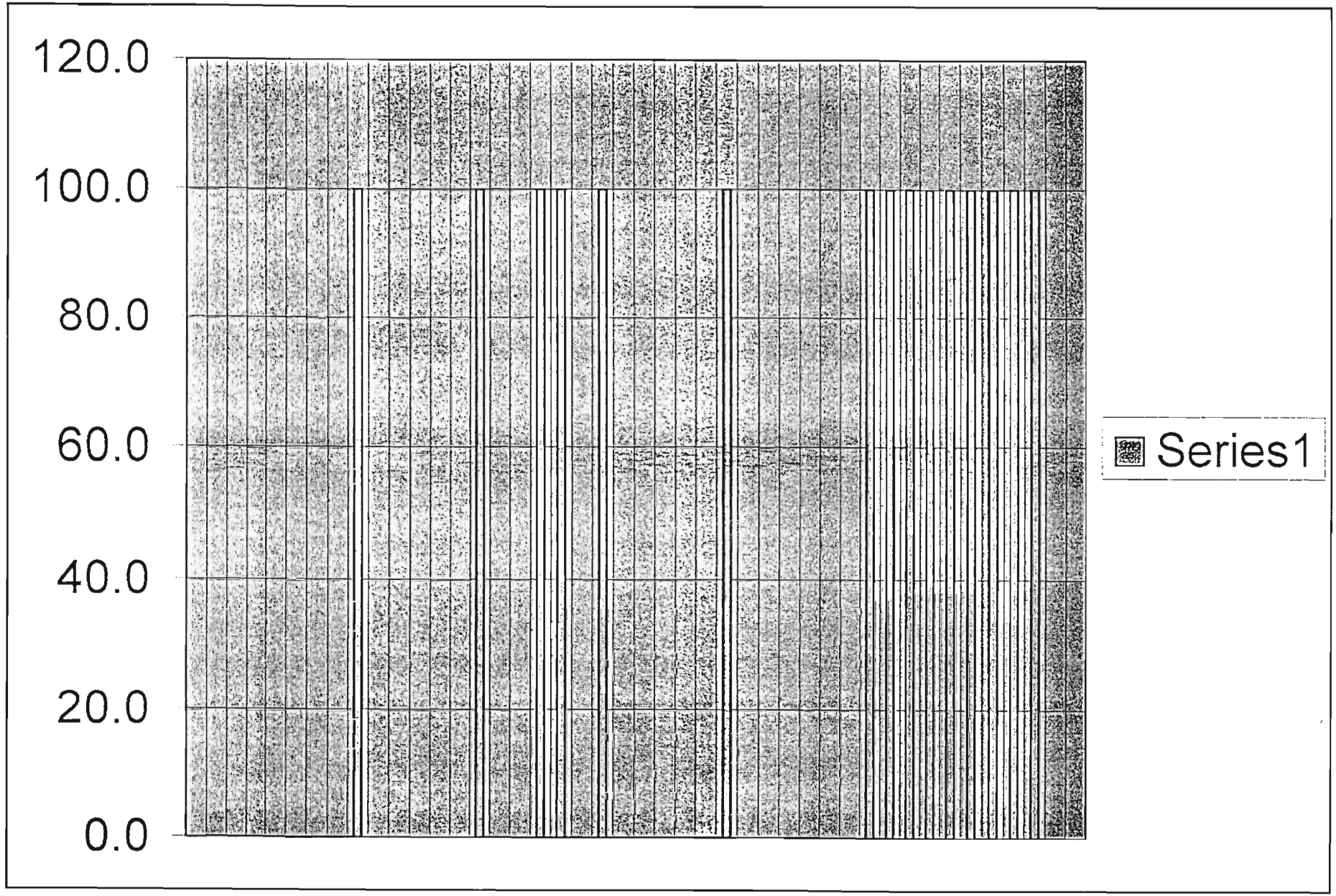
^{13}C NMR Data: (ppm)

175.4	146.2	130.9	125.2	117.6	86.2	52.9
52.5	51.3	47.0	43.3	38.3	36.2	35.9
34.1	33.6	33.6	33.3	32.0	29.8	28.2
27.9	27.3	25.7	25.0	23.0	18.3	18.3
18.3	13.3					



1

[8] Okorie, D.A. & Taylor, D.A.H., "Triterpenes from the seed of *Entandrophragma* species", *Phytochem.*, Pergamon Press, Great Britain, Vol. 16, pp2029-2030, 1977.

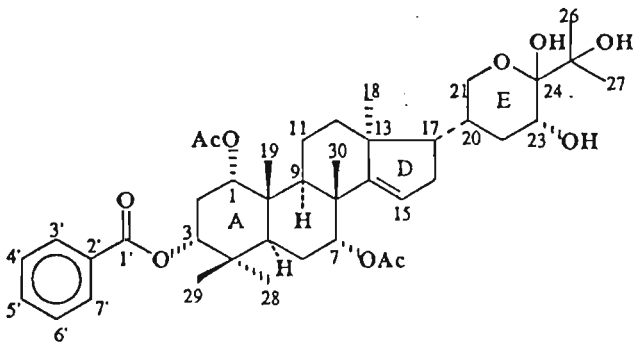


5 ppm Binned Spectrum of Entandrolide

Meliavolin [9]

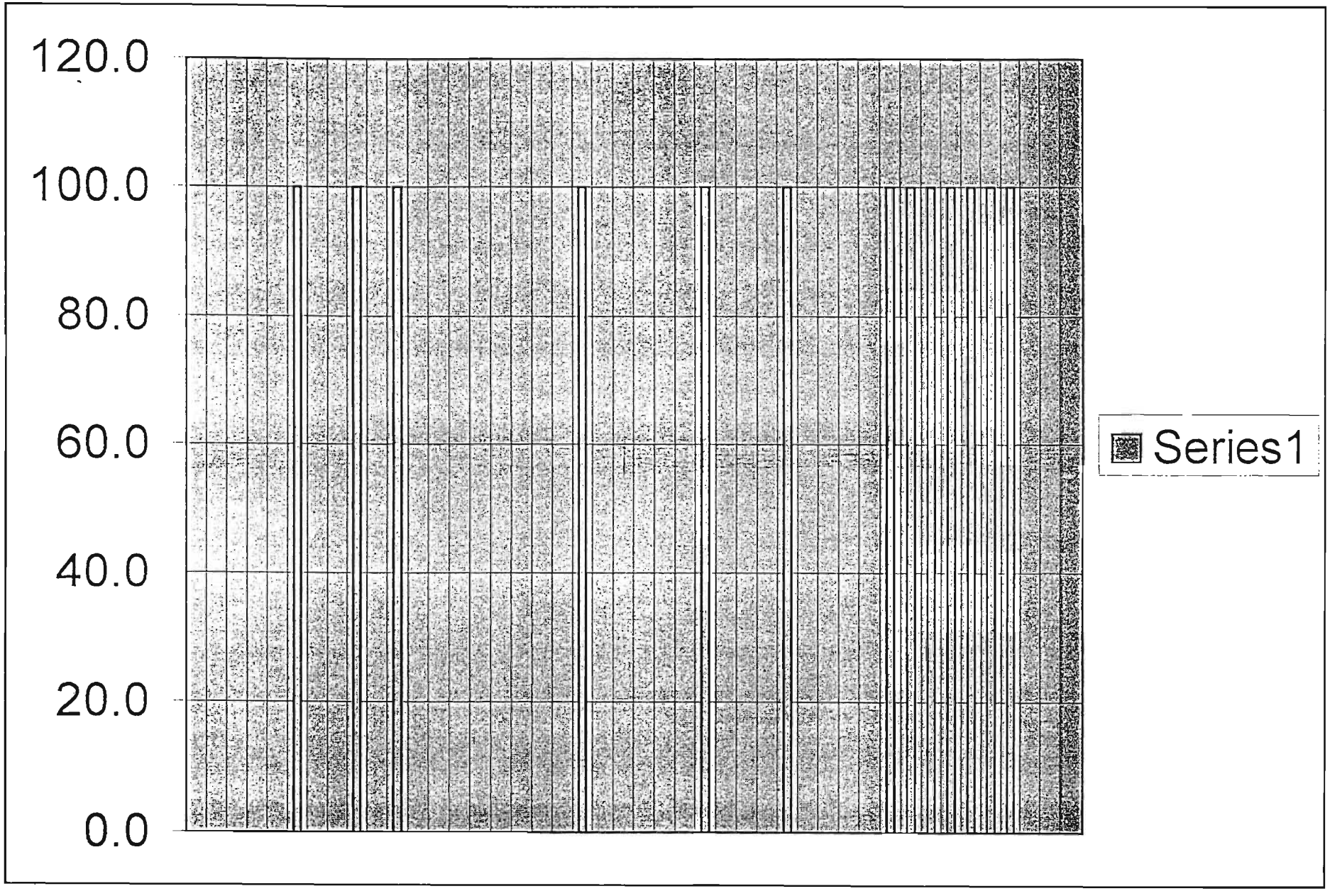
^{13}C NMR Data: (ppm)

170.1	169.7	165.2	158.9	133.0	130.6	129.4
129.4	128.2	128.2	119.3	95.7	77.1	76.3
75.5	72.7	67.5	65.4	56.9	46.4	42.0
40.2	37.4	36.5	35.3	34.7	33.8	32.7
29.2	28.0	26.9	25.4	24.1	23.2	22.8
21.4	21.0	20.9	20.0	16.1	16.0	



1

[9] Zeng, L.U., Gu, Z., Fang, X., Fanwick, P.E., Chang, C., Smith, D.L., & McLaughlin, J.L., "Two New Bioactive Triterpenoids from *Melia volkensii* (Meliaceae)", *Tetrahedron*, Elsevier Science, Great Britain, Vol. 51, no.9, pp2477-2488, 1995.

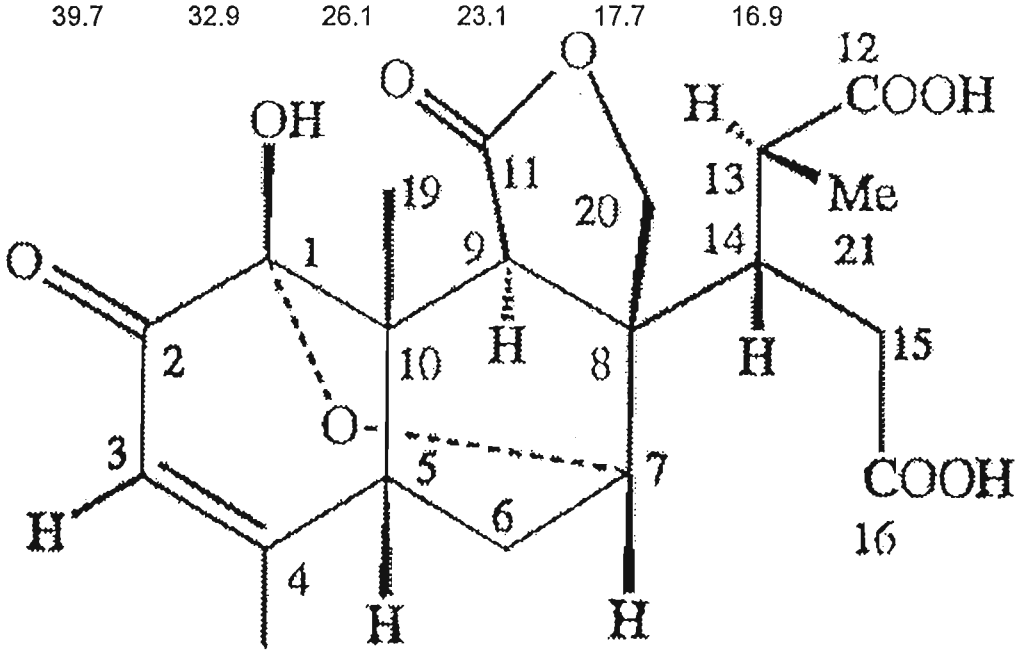


5 ppm Binned Spectrum of Meliavolin

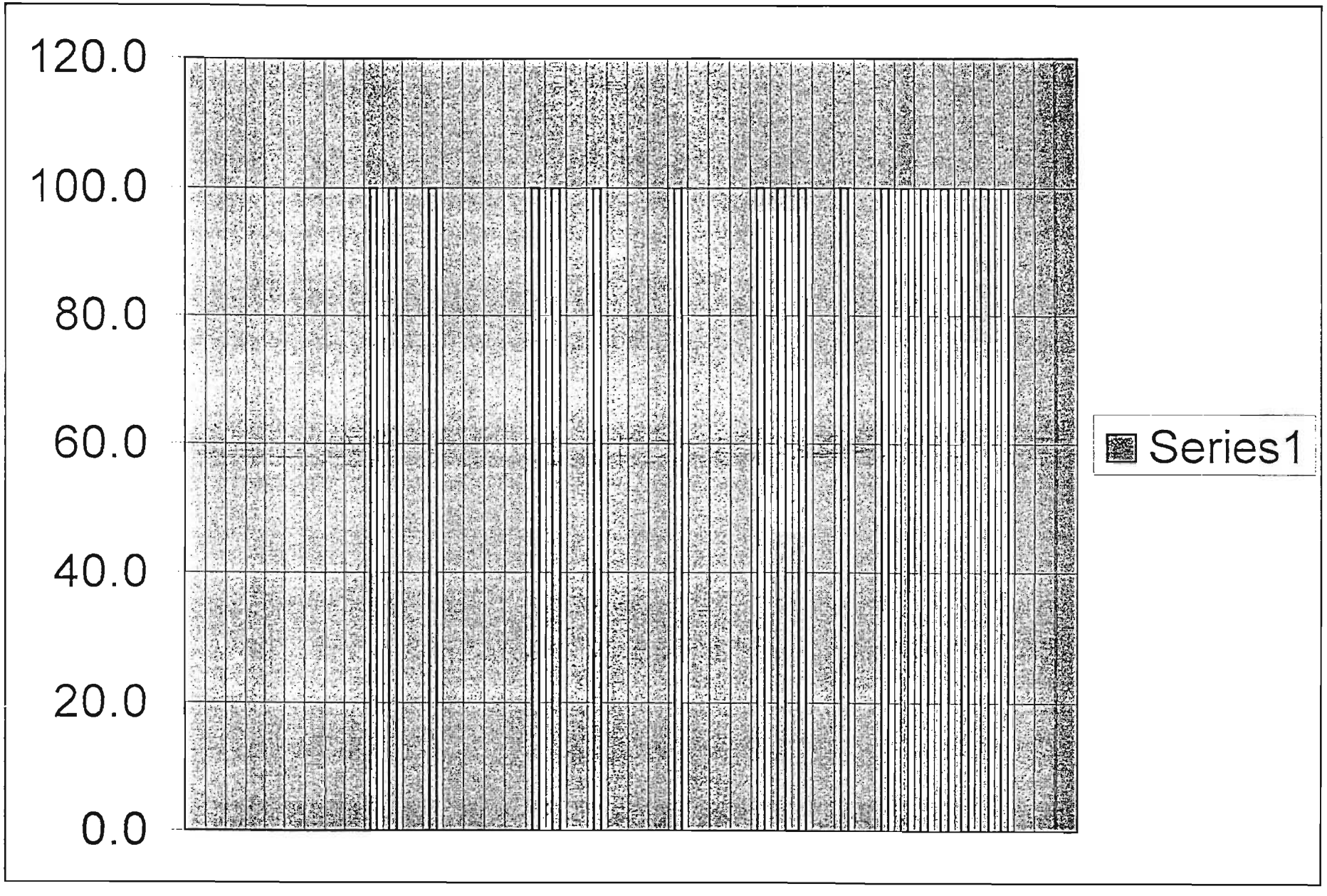
Vilmorinine A [11]

^{13}C NMR Data: (ppm)

192.6	178.5	177.5	176.2	166.4	120.1	94.0
72.1	70.5	49.3	46.6	42.2	40.9	39.9
39.7	32.9	26.1	23.1	17.7	16.9	



[11] Takeya,K., Kobata, H., Ozeki, A., Morita, H. & Itokawa, H., "A New Quassinoid from *Ailanthus vilmoriniana*", *J. Natl. Prod.*, ACS, Vol. 60(6), pp642-644, 1997.

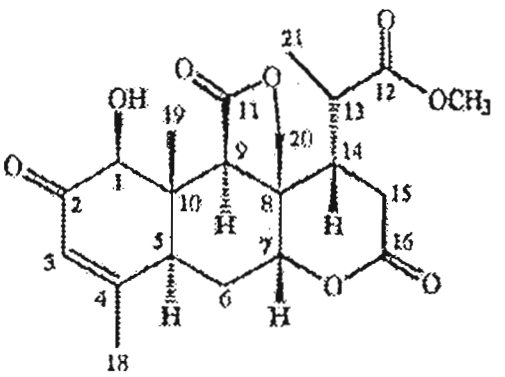


5 ppm Binned Spectrum of Vilmorinine A

Ailantinol A [12]

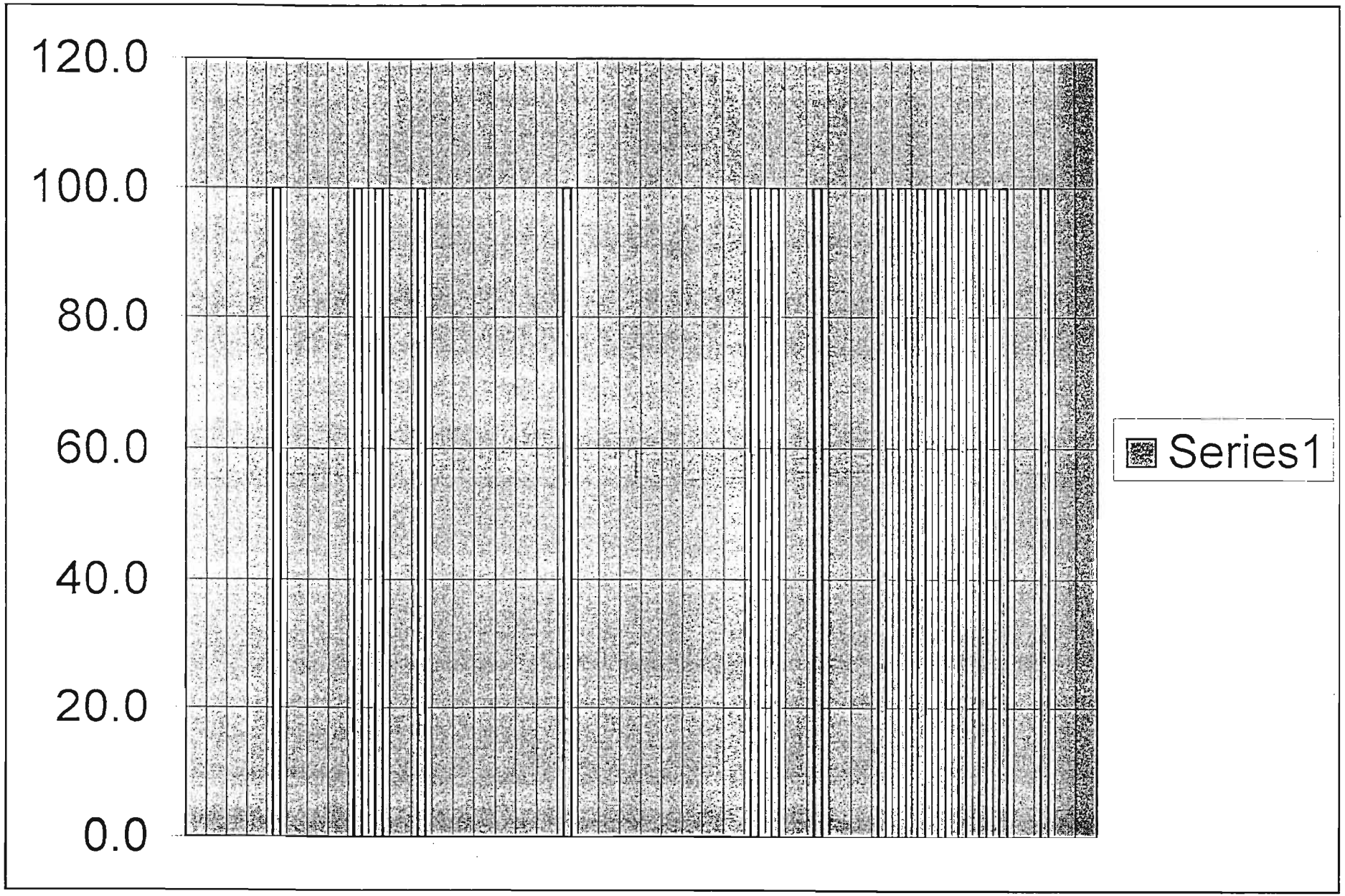
^{13}C NMR Data: (ppm)

197.5	175.6	172.8	172.3	160.3	126.9	84.7
75.7	69.4	54.3	52.2	45.8	45.0	40.9
40.6	36.9	31.4	26.0	22.2	10.7	10.7



1

[12] Kubota, K., Fukamiya, N., Hamada, T., Okano, M., Tagahara, K. & Lee, K-S., "Two New Quassinoids, Ailantinols A and B, and Related Compounds from *Ailanthus altissima*", *J. Natl. Prod.*, ACS, Vol. 59(7), pp683-686, 1996.

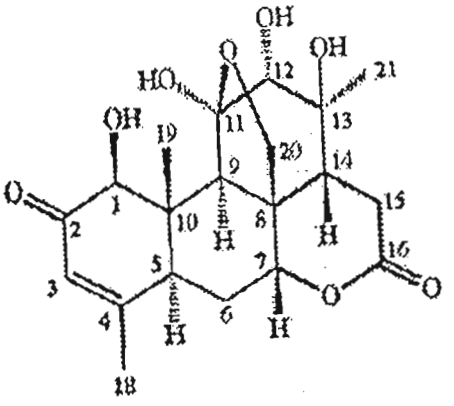


5 ppm Binned Spectrum of Ailantinol A

Ailantinol B [12]

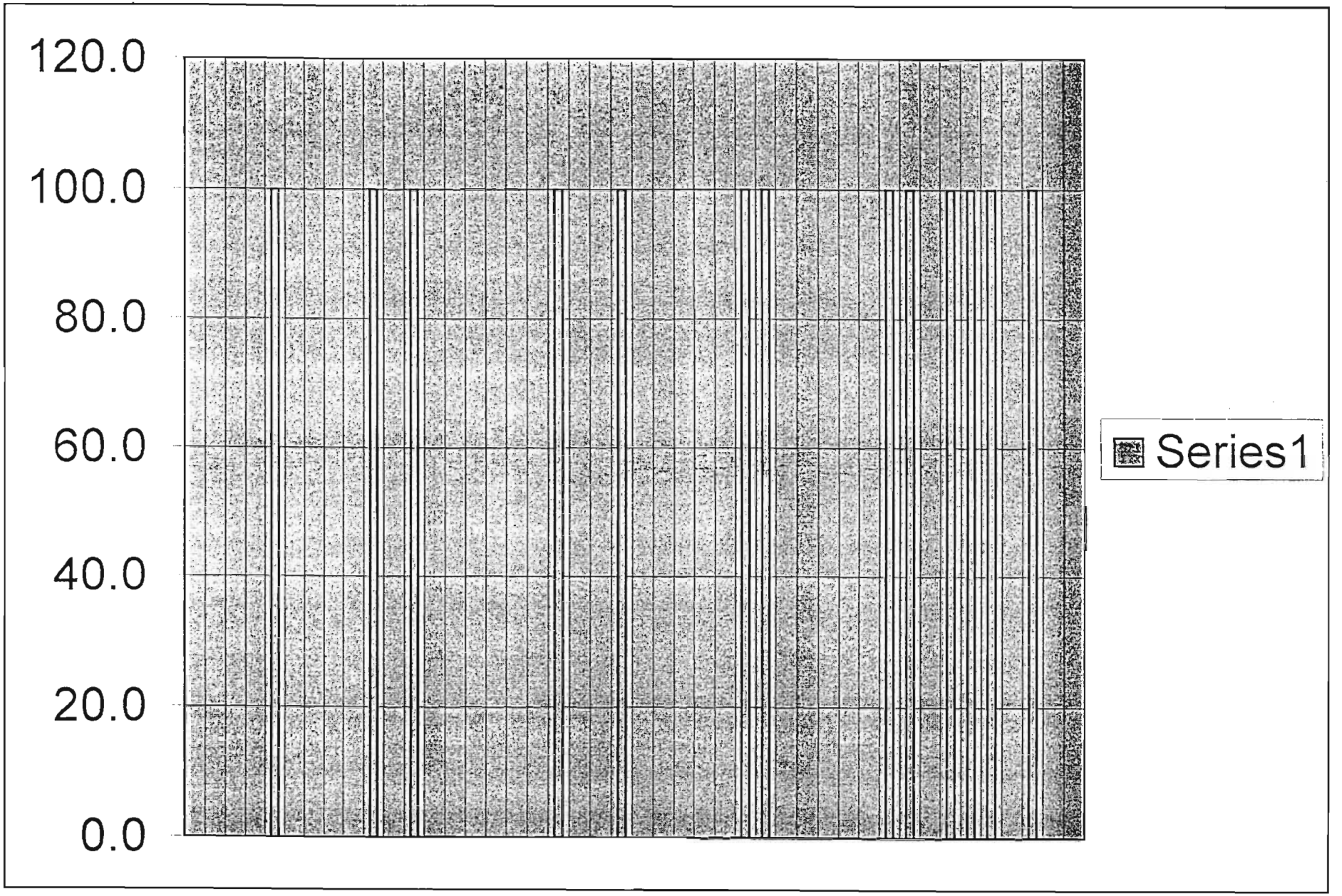
^{13}C NMR Data: (ppm)

197.6	170.2	162.3	126.3	110.7	84.6	83.0
78.2	74.2	71.0	49.0	46.5	45.3	44.8
42.6	31.8	26.2	26.1	22.4	10.7	



2

[12] Kubota, K., Fukamiya, N., Hamada, T., Okano, M., Tagahara, K. & Lee, K-S., "Two New Quassinoids, Ailantinols A and B, and Related Compounds from *Ailanthus altissima*", *J. Natl. Prod.*, ACS, Vol. 59(7), pp683-686, 1996.

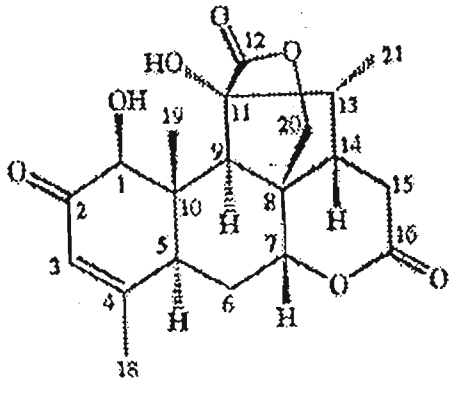


5 ppm Binned Spectrum of Ailantinol B

Shinjudilactone [12]

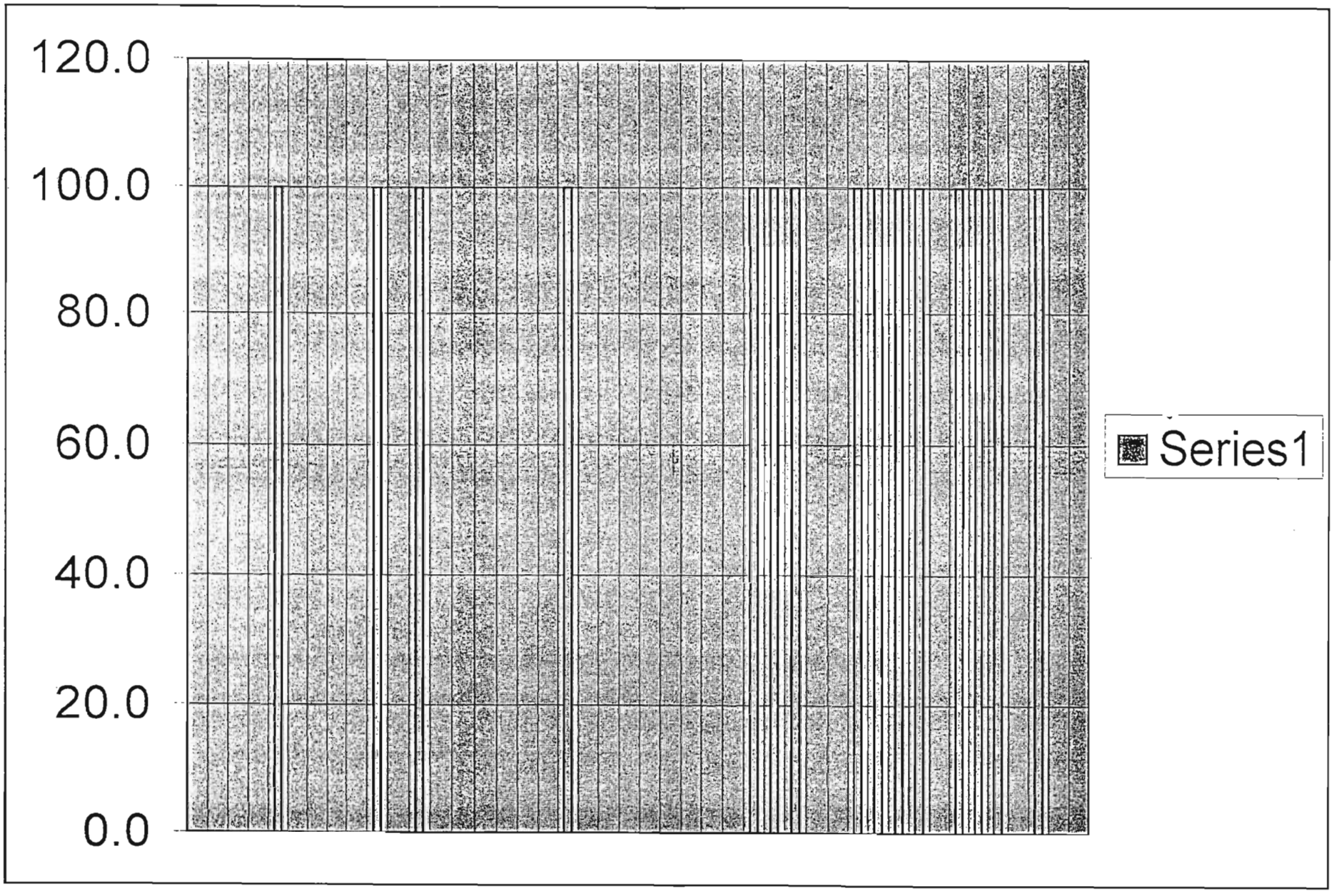
^{13}C NMR Data: (ppm)

196.9	173.5	170.6	162.0	126.3	83.8	78.7
76.2	73.9	55.0	53.6	48.4	45.6	43.0
42.2	32.9	27.0	22.1	13.8	10.6	



3

[12] Kubota, K., Fukamiya, N., Hamada, T., Okano, M., Tagahara, K. & Lee, K-S., "Two New Quassinoids, Ailantinols A and B, and Related Compounds from *Ailanthus altissima*", *J. Natl. Prod.*, ACS, Vol. 59(7), pp683-686, 1996.

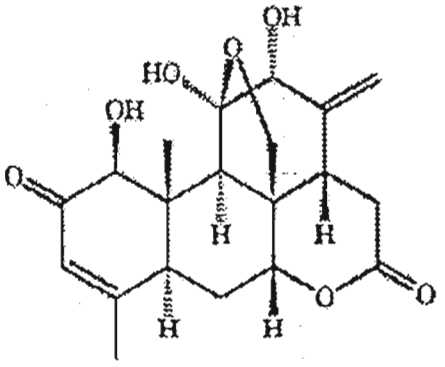


5 ppm Binned Spectrum of Shinjudilactone

Ailanthone [12]

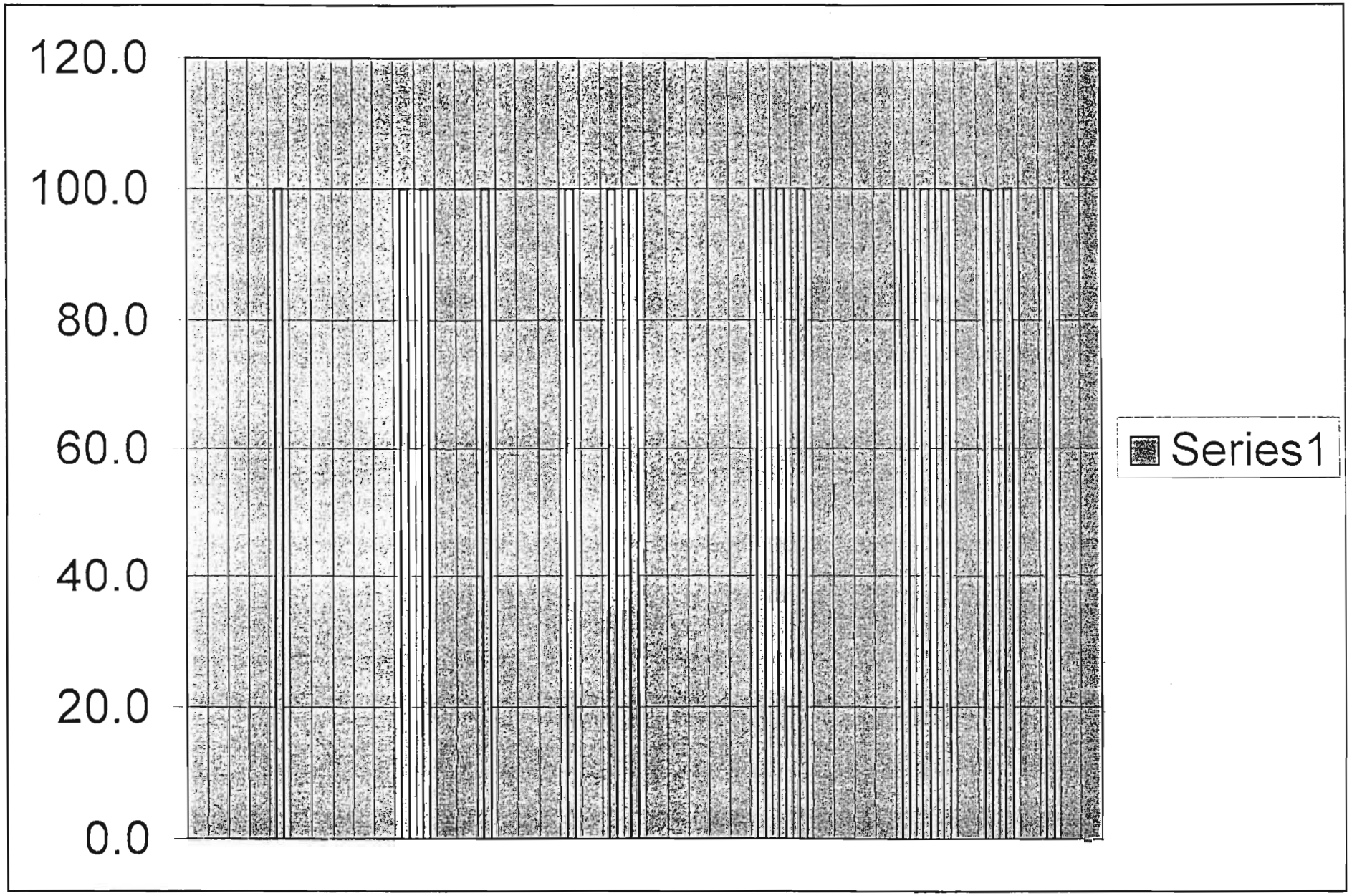
^{13}C NMR Data: (ppm)

197.3	169.4	162.1	147.4	126.2	118.2	110.3
84.4	80.6	78.5	72.3	48.0	45.7	45.5
44.8	42.5	35.3	26.2	22.5	10.3	



4

[12] Kubota, K., Fukamiya, N., Hamada, T., Okano, M., Tagahara, K. & Lee, K-S., "Two New Quassinoids, Ailantinols A and B, and Related Compounds from *Ailanthus altissima*", *J. Natl. Prod.*, ACS, Vol. 59(7), pp683-686, 1996.

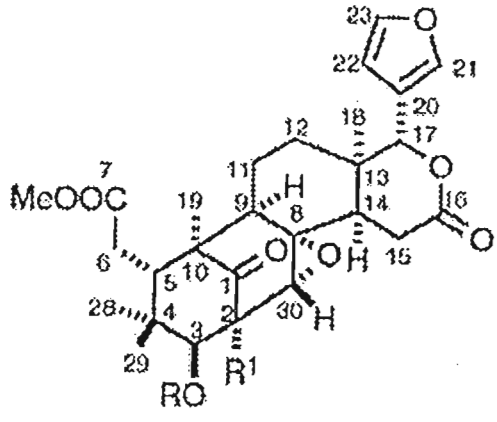


5 ppm Binned Spectrum of Ailanthone

Xylocarpin [14]

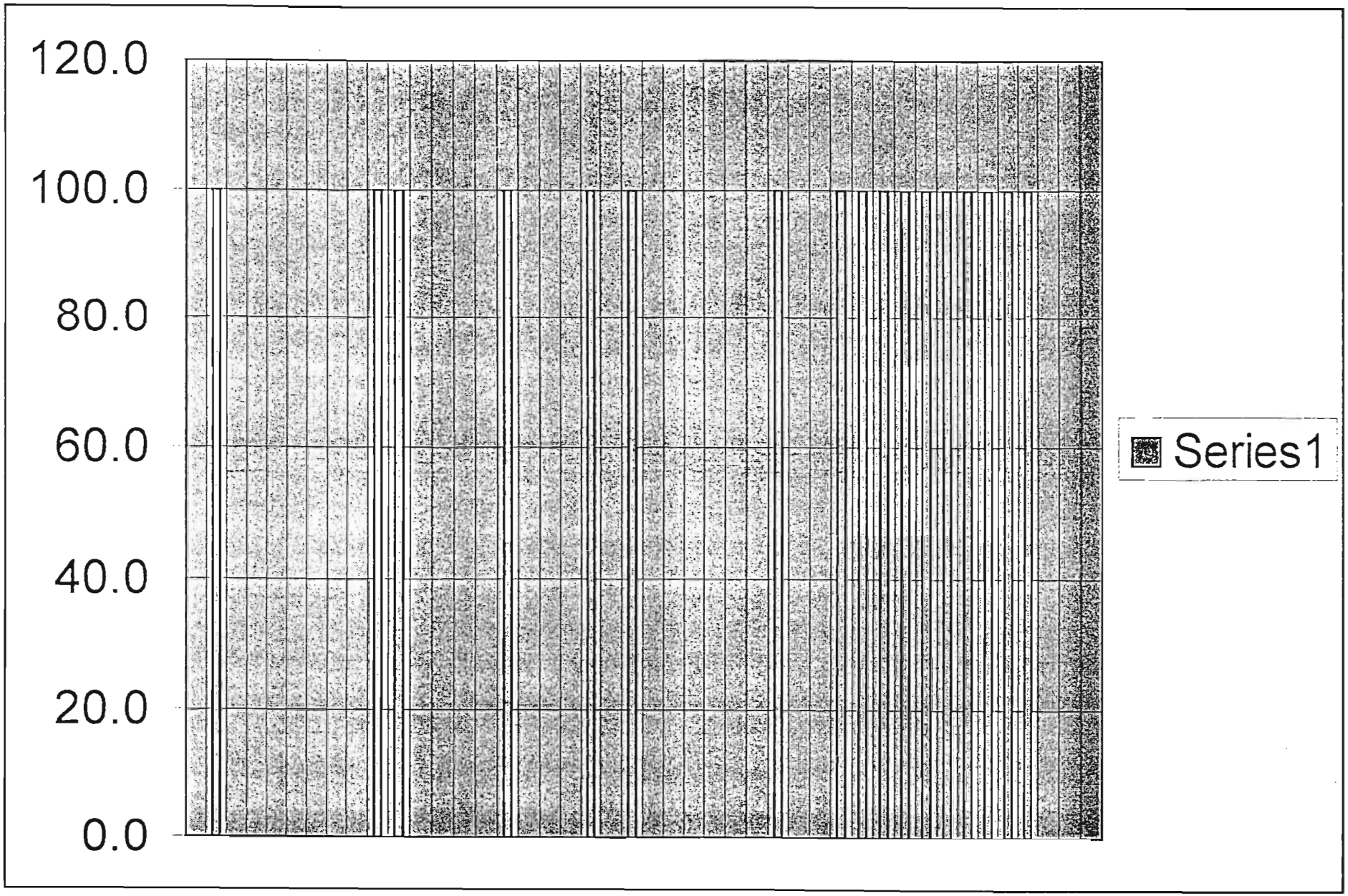
^{13}C NMR Data: (ppm)

214.0	174.2	172.1	169.8	143.1	141.0	120.1	110.2
78.8	77.2	63.6	60.5	55.8	52.3	48.6	48.2
45.9	42.5	39.2	36.4	33.7	33.4	33.1	26.3
22.5	20.8	20.7	19.4	16.0			



2: R = Ac, R^1 = H

[14] Mootoo, B.S., Ramsewak, R., Khan, A., Tinto, W.F., Reynolds, W.F., McLean, S. & Yu, M., "Tetranortriterpenoids from *Ruagea glabra*", *J. Natl. Prod.*, ACS, Vol. 59(5), pp544-547, 1996.

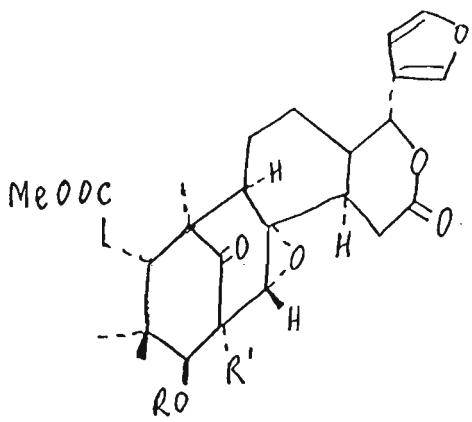


5 ppm Binned Spectrum of Xylocarpin

Ruageanin B [14]

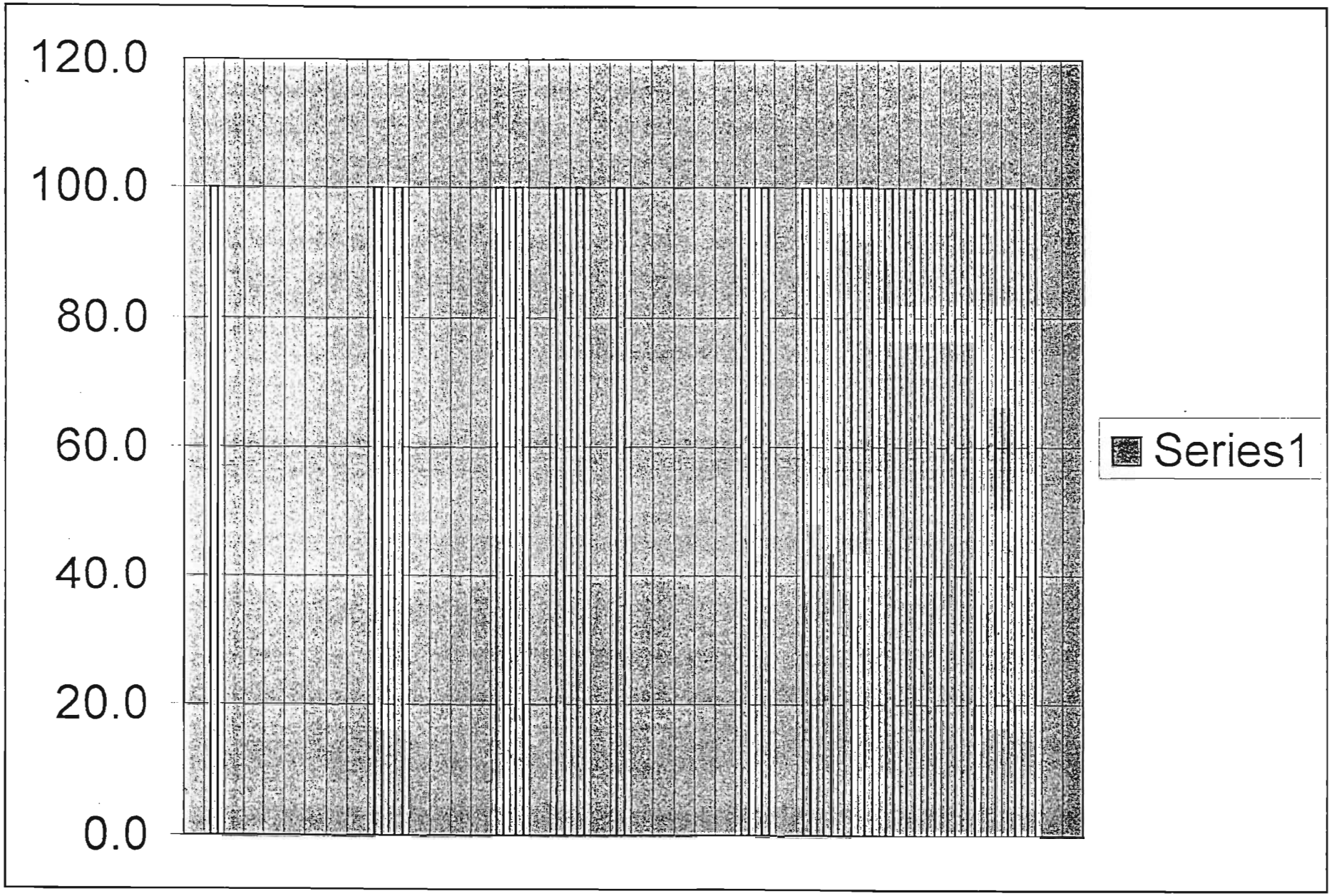
^{13}C NMR Data: (ppm)

213.0	173.9	171.2	166.9	143.1	140.9	139.8	127.8
120.2	110.1	84.8	78.9	78.4	67.4	63.1	55.1
52.4	49.1	45.2	42.3	40.1	36.2	33.5	33.2
32.9	26.3	22.0	20.5	19.4	16.1	14.6	12.6



4: R = $COCCH_3$=$CHCH_3$, R^1 = OH

[14] Mootoo, B.S., Ramsewak, R., Khan, A., Tinto, W.F., Reynolds, W.F., McLean, S. & Yu, M., "Tetranortriterpenoids from *Ruagea glabra*", *J. Natl. Prod.*, ACS, Vol. 59(5), pp544-547, 1996.

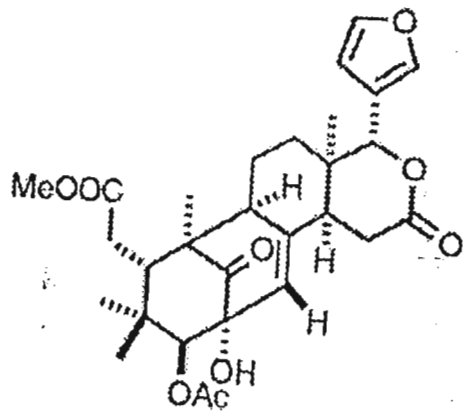


5 ppm Binned Spectrum of Ruageanin B

Ruageanin D [14]

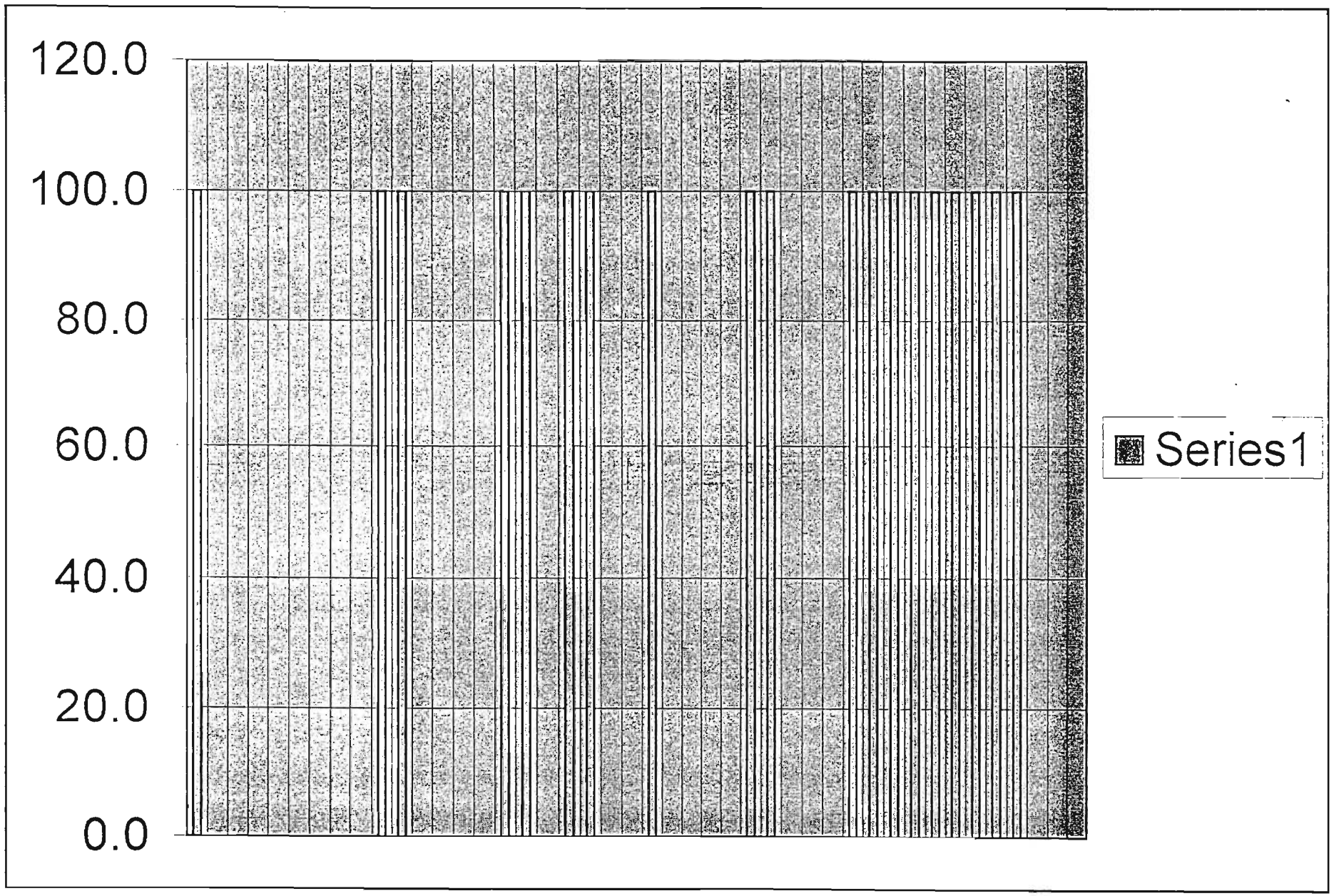
^{13}C NMR Data: (ppm)

215.4	174.3	171.2	169.3	142.9	141.9	136.5
129.1	120.4	109.7	84.8	77.1	76.7	56.3
52.2	48.8	44.8	41.0	38.4	36.1	33.7
32.6	29.9	21.9	21.8	20.3	19.5	19.1
15.7						



6

[14] Móotoo, B.S., Ramsewak, R., Khan, A., Tinto, W.F., Reynolds, W.F., McLean, S. & Yu, M., "Tetranortriterpenoids from *Ruagea glabra*", *J. Natl. Prod.*, ACS, Vol. 59(5), pp544-547, 1996.

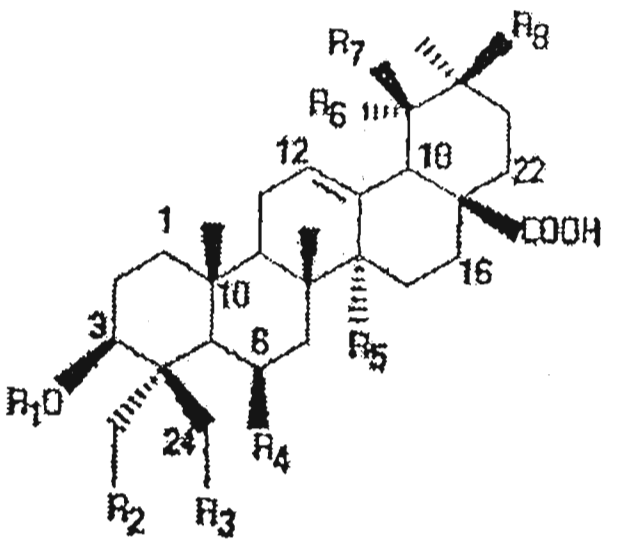


5 ppm Binned Spectrum of Ruageanin D

3β,23,24-trihydroxyolean-12-en-28-oic acid [13]

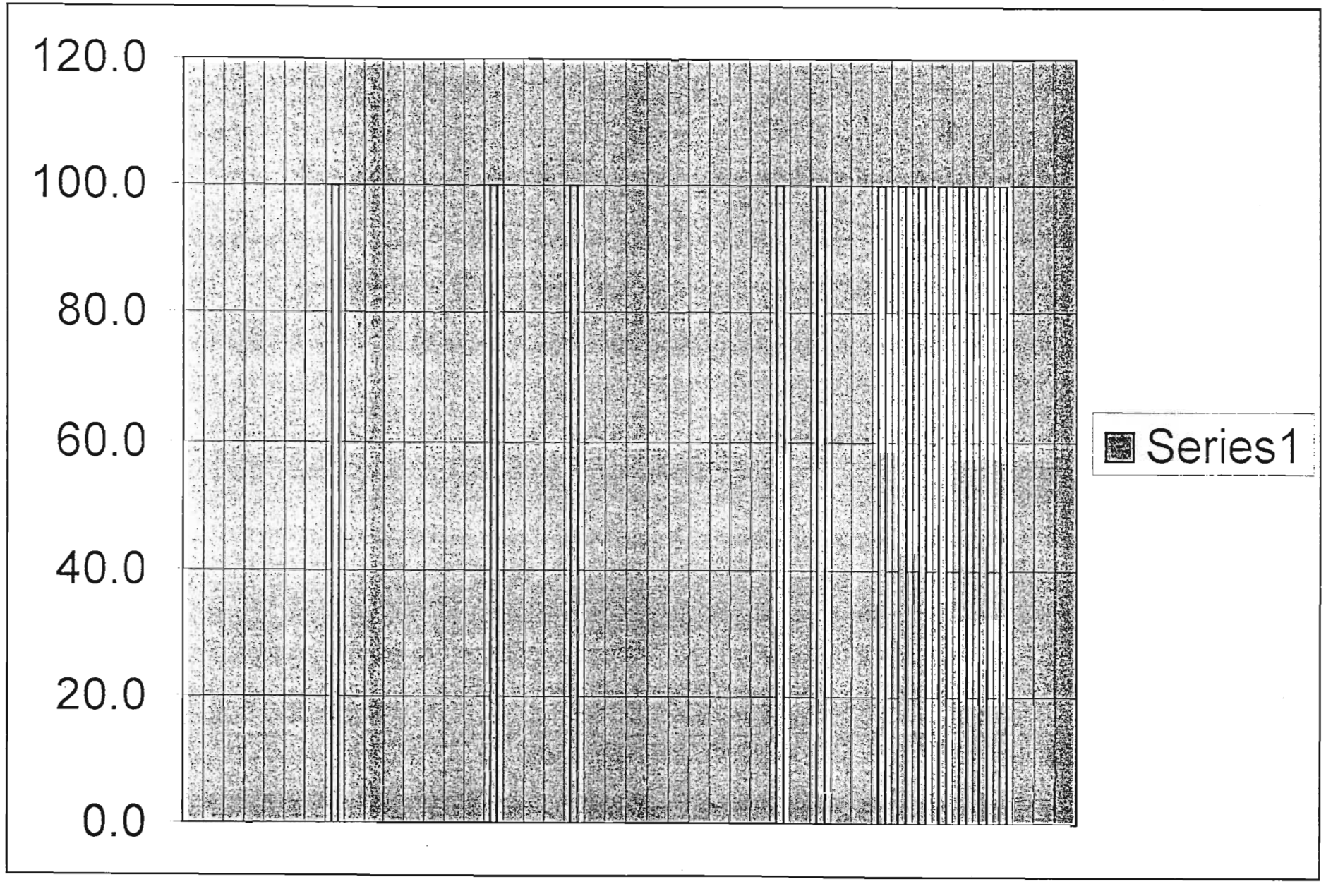
^{13}C NMR Data: (ppm)

180.0	144.6	122.3	74.1	63.1	63.1	48.3
48.1	46.8	46.5	46.3	41.9	41.8	39.6
38.6	36.8	34.0	33.2	33.0	33.0	30.8
28.1	28.0	25.9	23.9	23.6	23.5	19.0
17.1	15.7					



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
1	H	OH	OH	H	Me	H	H	Me

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

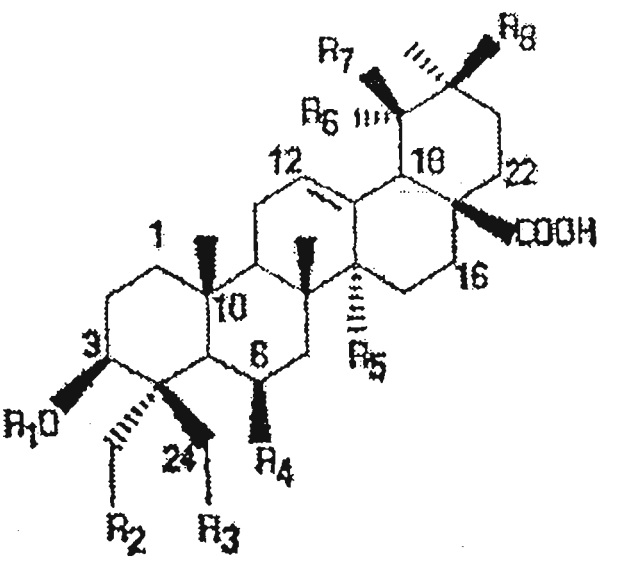


5 ppm Binned Spectrum of 3β,23,24-trihydroxyolean-12-en-28-oic acid

3β,6β,24-trihydroxyolean-12-en-28-oic acid [13]

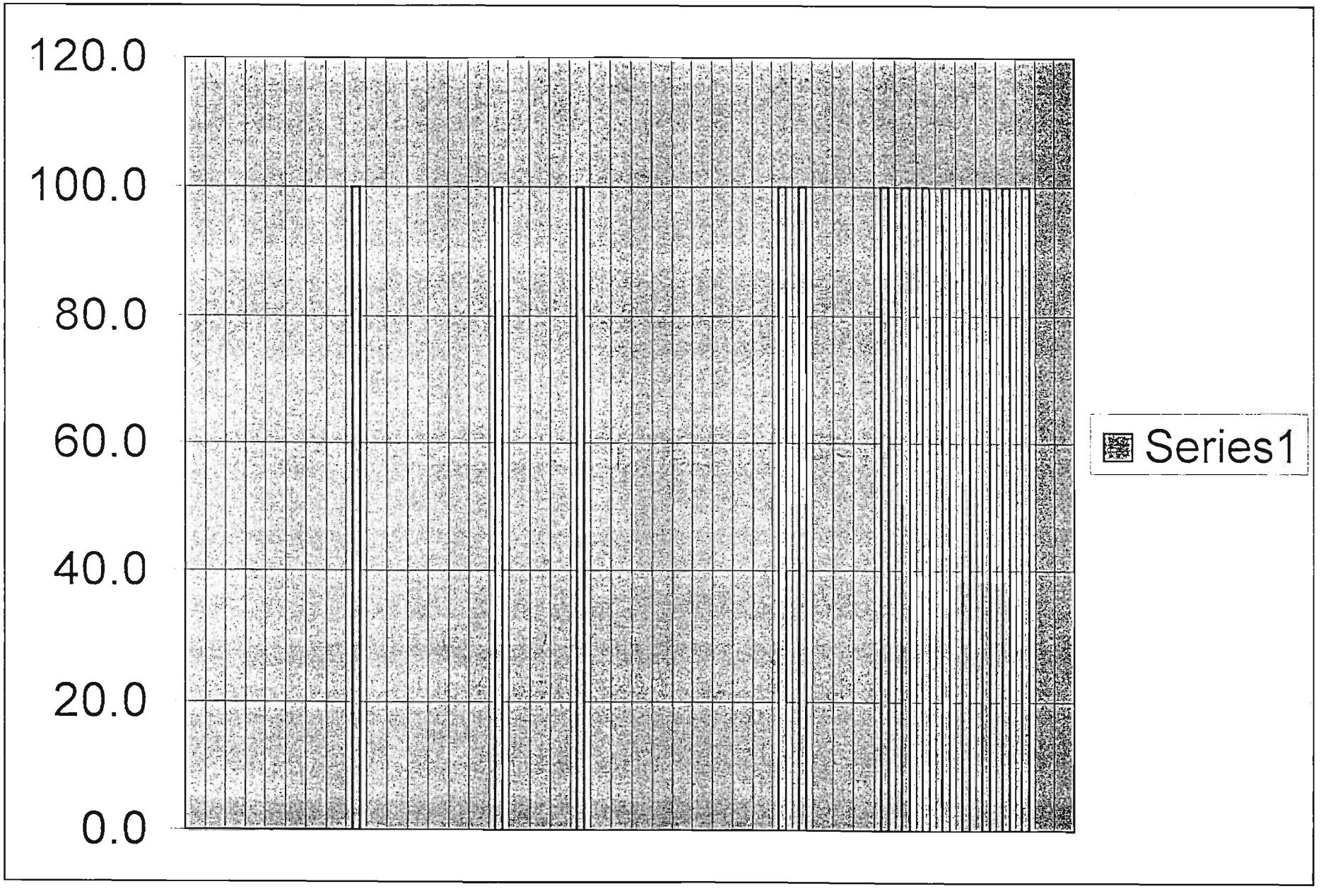
^{13}C NMR Data: (ppm)

179.5	143.5	122.3	72.6	66.9	66.4	48.7
48.0	46.0	45.8	43.4	42.0	41.4	40.4
40.1	38.5	36.3	33.5	32.6	32.6	30.3
27.6	27.3	25.6	23.2	23.1	23.1	17.9
16.8	14.1					



	R_1	R_3	R_3	R_4	R_5	R_6	R_7	R_8
2	H	H	OH	OH	Me	H	H	Me

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

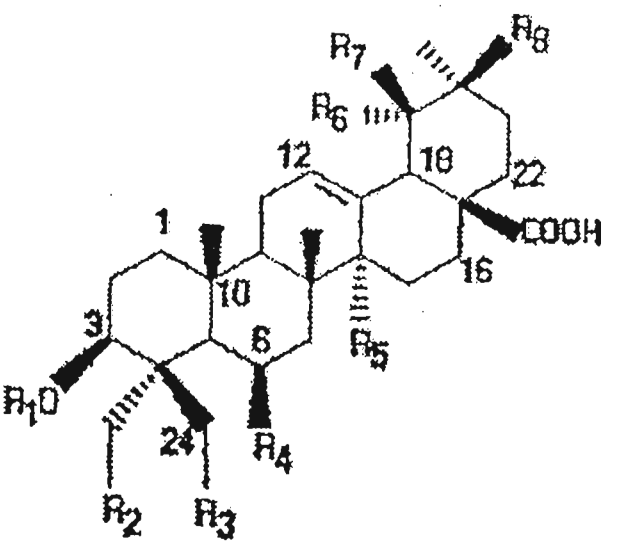


5 ppm Binned Spectrum of 3β,6β,24-trihydroxyolean-

3β,6β,19α,24-tetrahydroxyurs-12-en-28-oic acid [13]

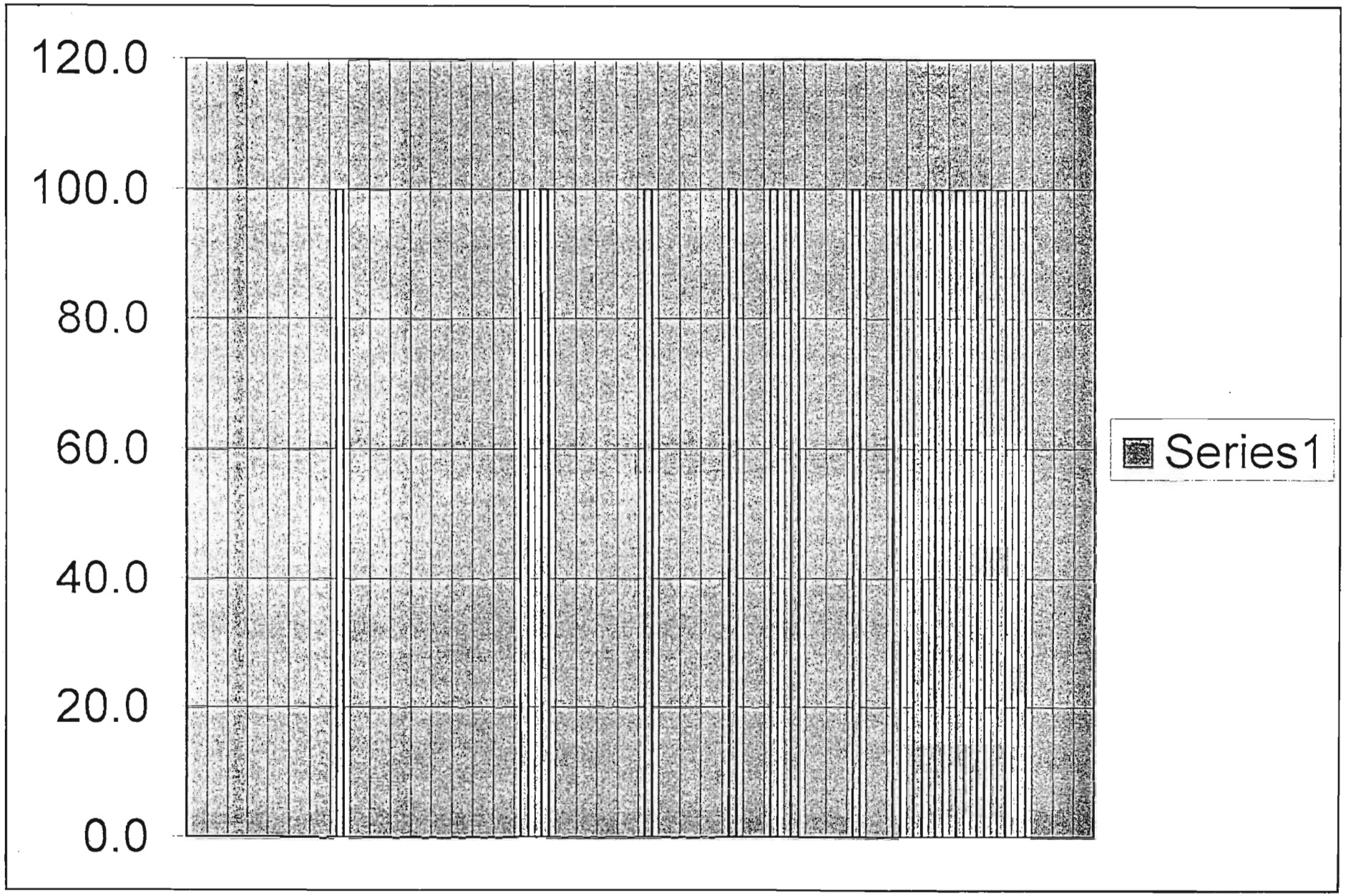
^{13}C NMR Data: (ppm)

180.9	139.5	128.5	73.9	73.0	68.0	68.0	54.9
49.9	48.5	48.5	44.0	42.8	42.5	41.6	41.3
39.9	38.6	37.1	29.4	28.1	27.3	27.1	26.7
24.8	24.4	18.5	17.6	16.8	14.6		



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
3	H	H	OH	OH	Me	OH	Me	H

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.



5 ppm Binned Spectrum of Pyrocincholic acid 3β-O-β-D-fucopyranoside

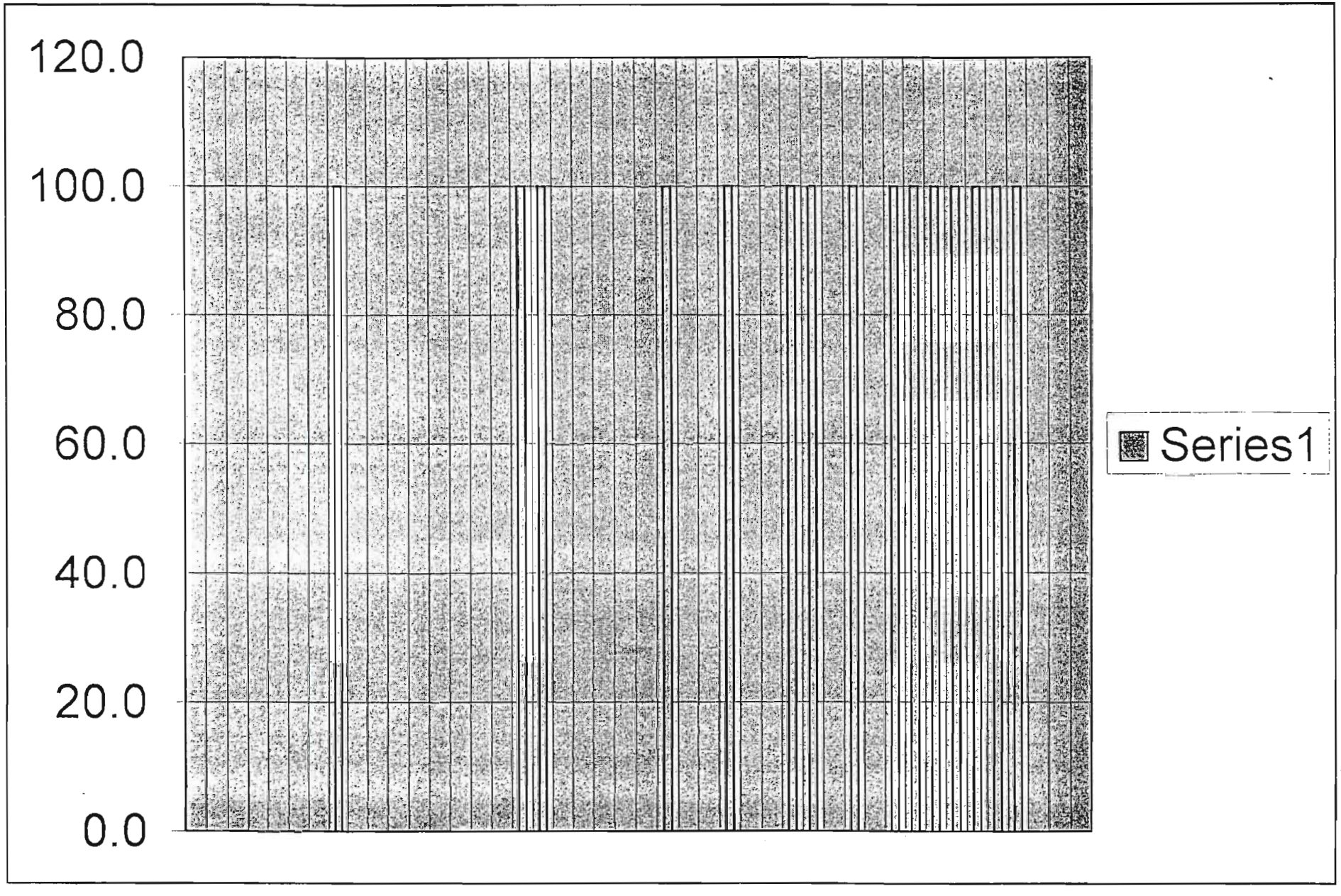
Pyrocincholic acid 3β-O-α-L-rhamnopyranoside [13]

^{13}C NMR Data: (ppm)

180.2	136.8	130.7	104.5	88.8	74.1	73.0
72.5	69.9	56.3	55.4	45.2	41.7	39.8
39.5	39.3	38.2	37.9	37.2	34.6	32.5
32.1	31.7	30.8	28.2	26.0	25.1	24.5
21.2	20.8	18.8	18.5	18.1	16.5	16.5

STRUCTURE NOT GIVEN IN PAPER

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

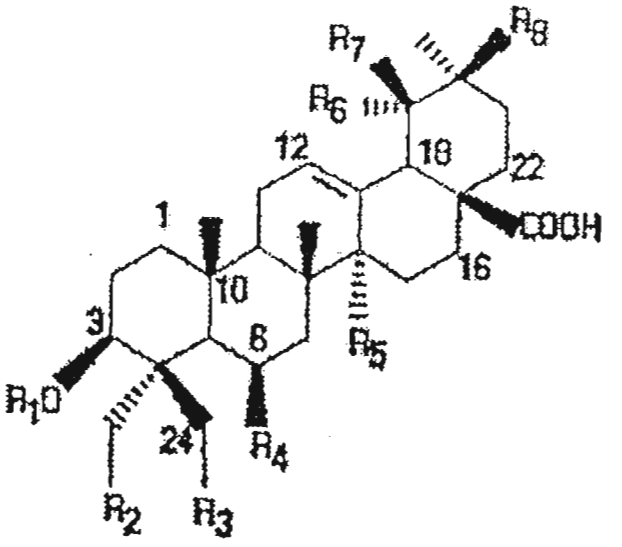


5 ppm Binned Spectrum of Pyrocincholic acid 3β-O-α-L-rhamnopyranoside

Compound 7 [13]

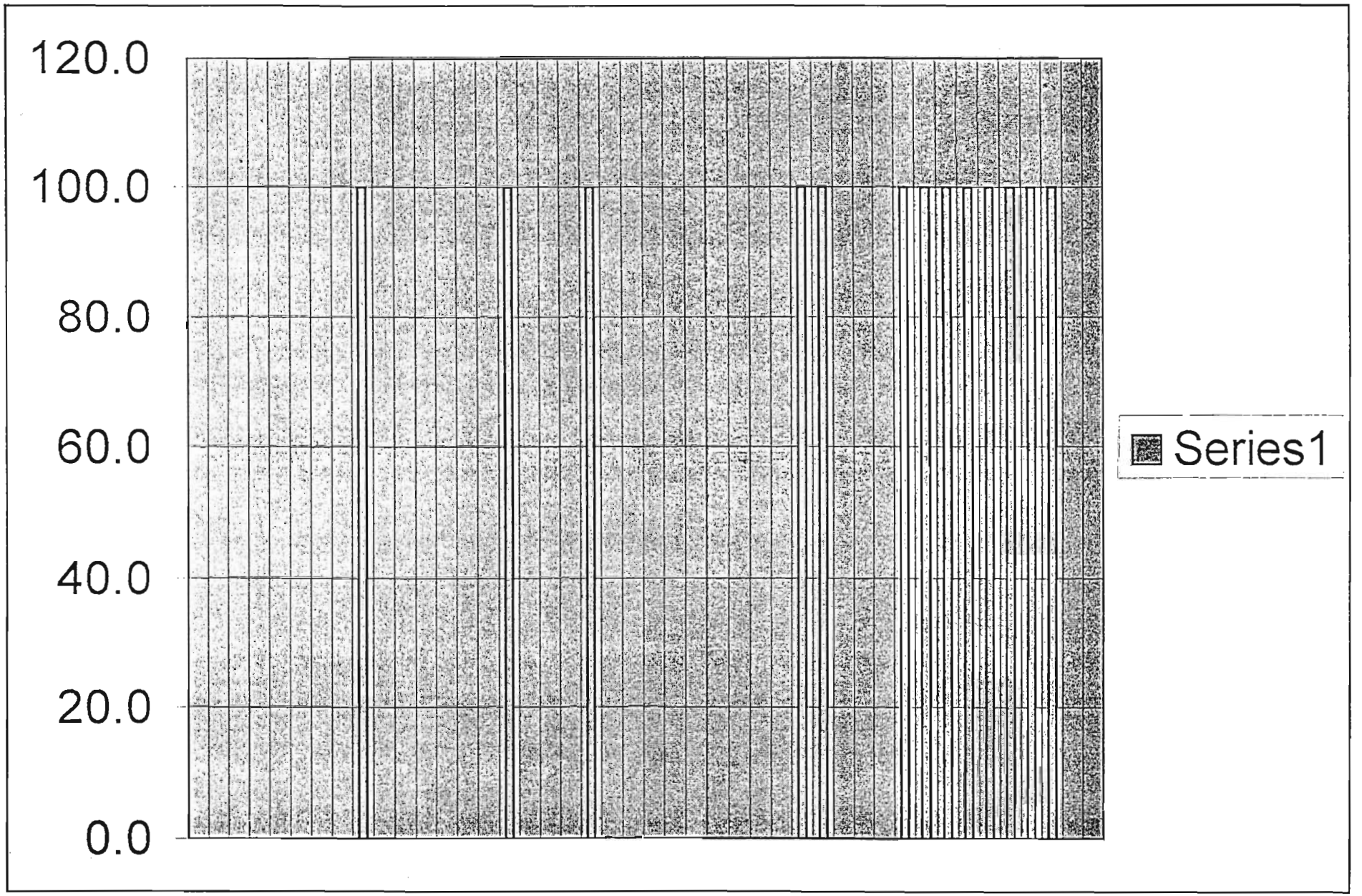
^{13}C NMR Data: (ppm)

179.9	144.6	122.2	73.3	67.9	48.5	47.9
46.4	46.2	42.6	42.0	41.8	39.5	38.5
37.0	34.0	33.0	33.0	32.7	30.7	28.1
27.4	25.9	23.6	23.5	23.5	18.3	17.2
15.7	12.8					



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
7	H	OH	H	H	Me	H	H	Me

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

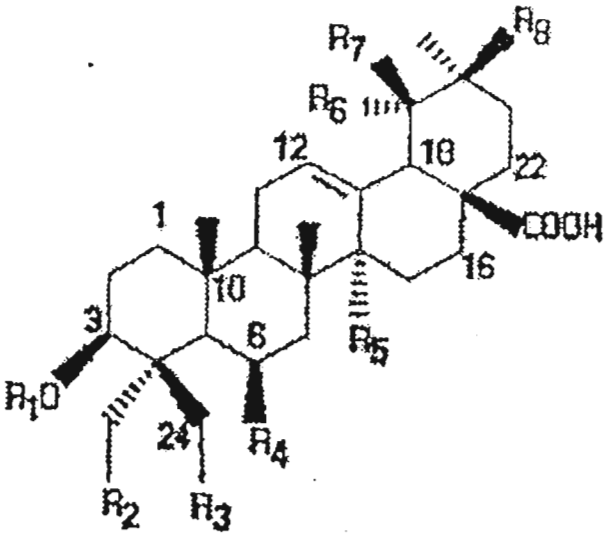


5 ppm Binned Spectrum of Compound 7

Compound 8[13]

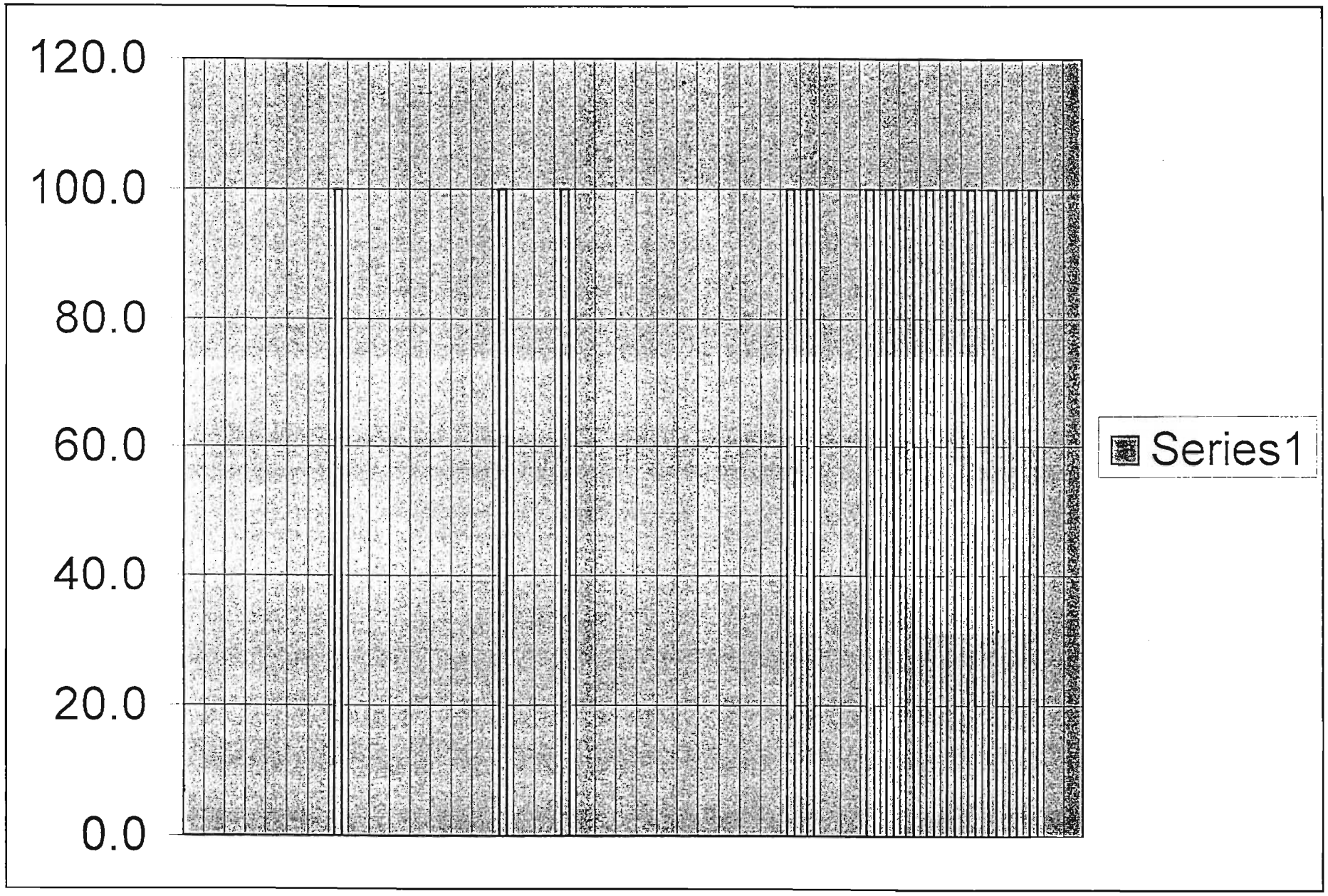
^{13}C NMR Data: (ppm)

180.7	140.0	128.1	73.7	72.7	68.2	54.7
48.8	48.3	47.9	42.9	42.4	42.2	40.4
38.9	38.5	37.3	33.4	29.4	27.7	27.2
27.0	26.5	24.9	24.1	18.9	17.3	16.8
16.0	13.1					



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
8	H	OH	H	H	Me	OH	Me	H

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

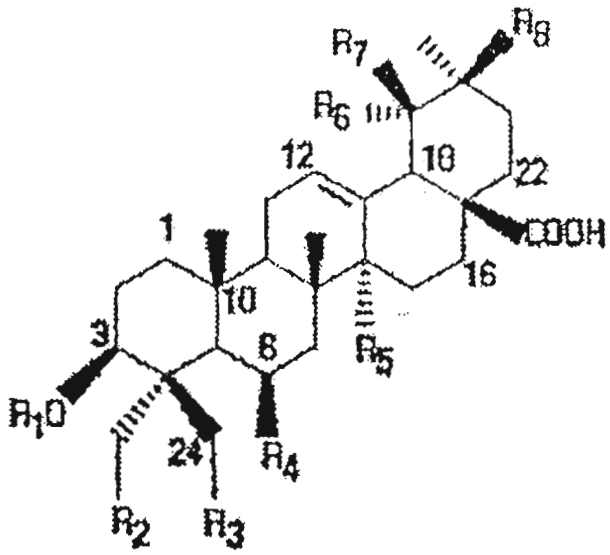


5 ppm Binned Spectrum of Compound 8

Compound 9 [13]

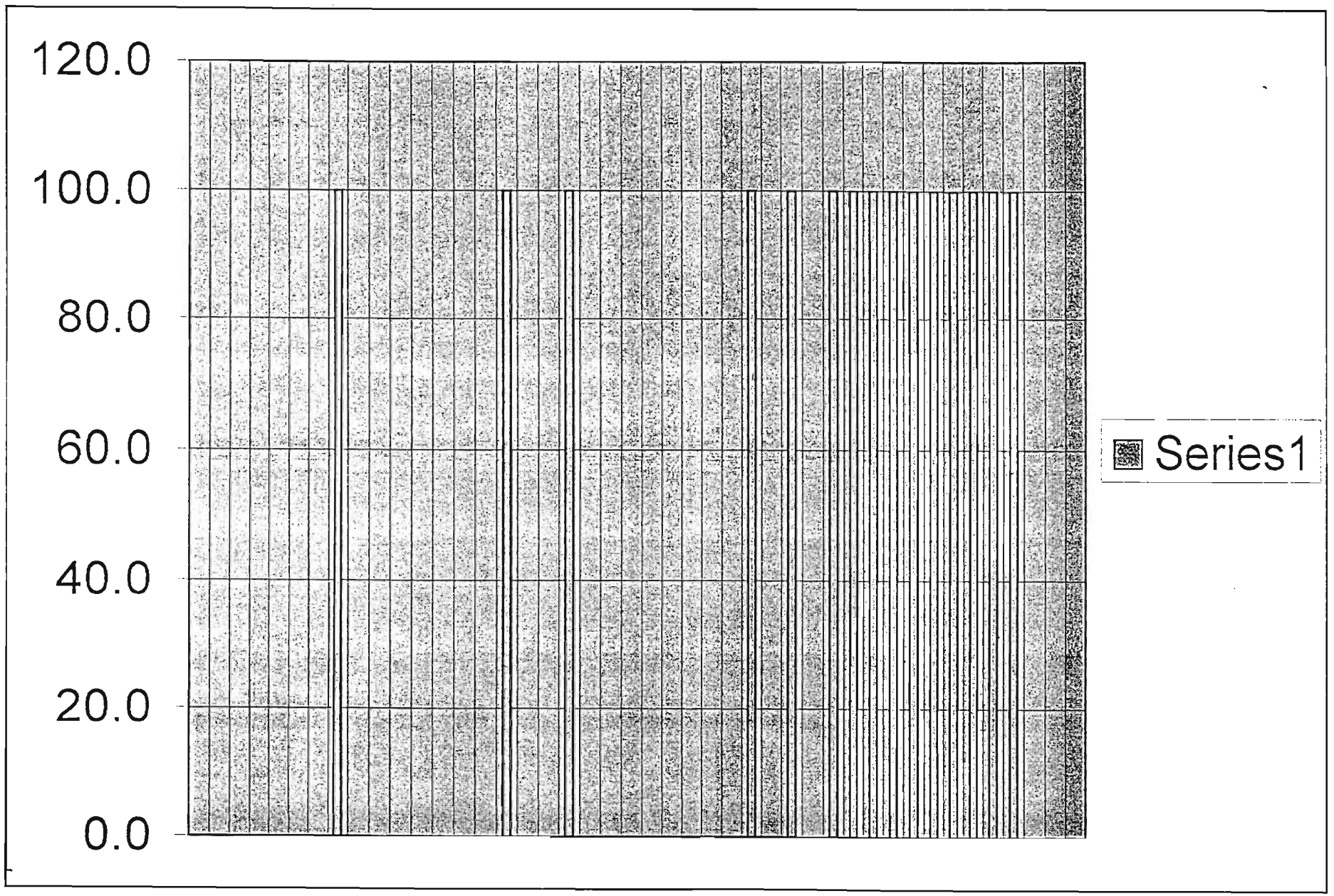
^{13}C NMR Data: (ppm)

180.8	140.0	127.9	80.3	72.8	64.6	56.5	54.7
48.4	47.9	43.2	42.4	42.1	40.4	38.8	38.6
37.2	34.0	29.1	28.5	27.2	27.0	26.5	24.7
24.3	23.7	19.3	17.2	16.8	16.1		



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
9	H	H	OH	H	Me	OH	Me	H

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

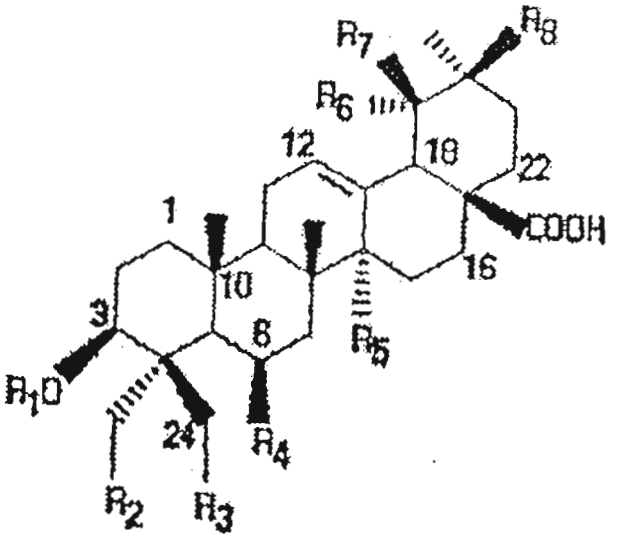


5 ppm Binned Spectrum of Compound 9

Compound 10 [13]

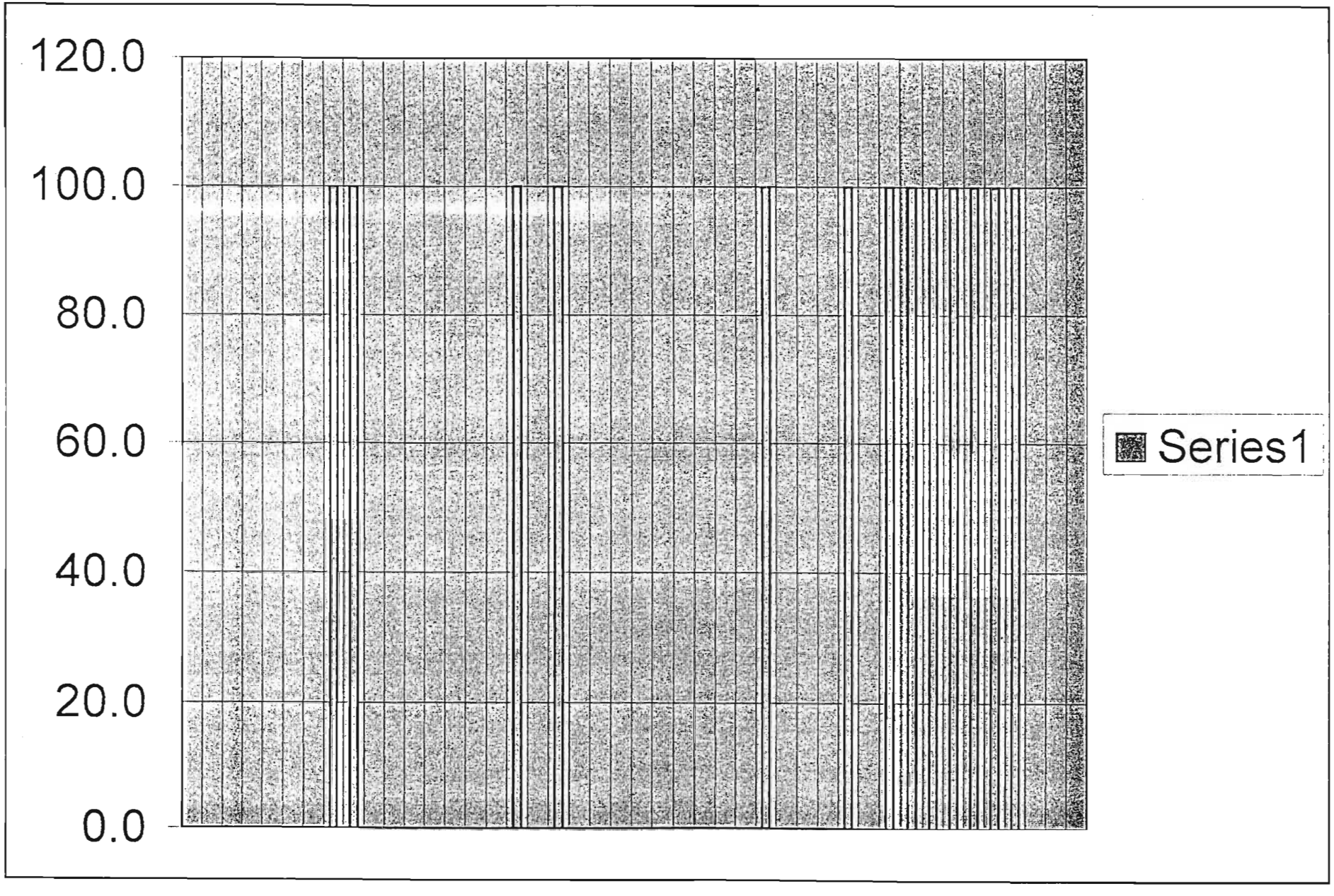
^{13}C NMR Data: (ppm)

180.2	178.5	138.1	125.9	77.8	56.5	55.7	47.7
47.5	44.2	44.0	39.8	39.2	39.1	37.4	37.3
34.0	33.1	32.8	30.9	28.5	28.1	25.3	24.9
23.7	23.4	18.7	18.7	16.5	16.4		



	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
10	H	H	H	H	COOH	H	H	Me

[13] Fang, S-Y., He, Z-S. & Fan G-J., "Triterpenoids from *Adina rubella*", *J. Natl. Prod.*, ACS, Vol. 59(3), pp304-307, 1996.

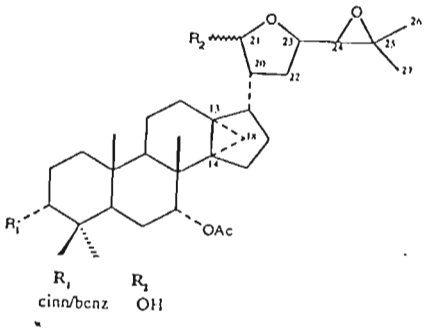


5 ppm Binned Spectrum of Compound 10

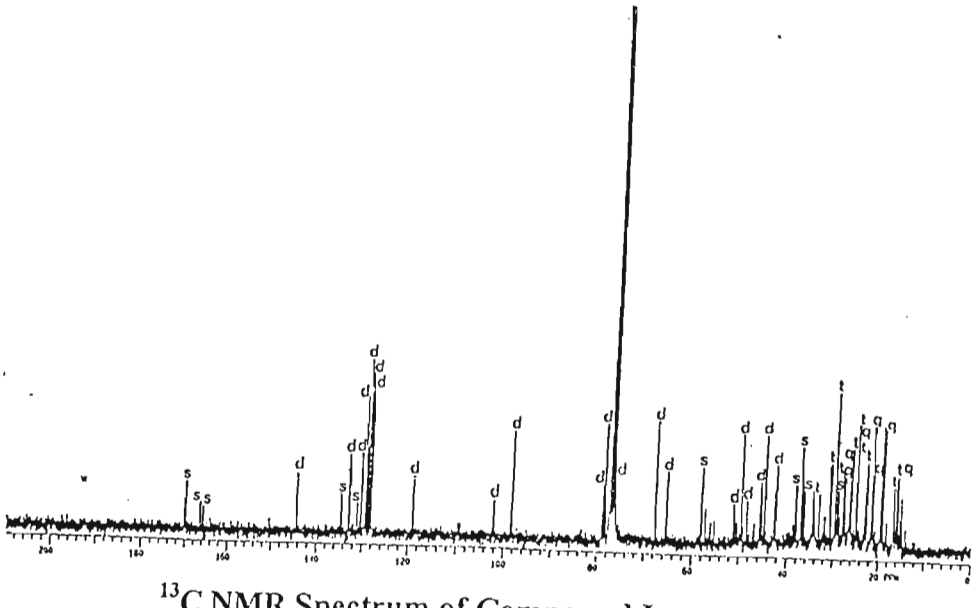
^{13}C NMR Data for Impure Compound L

ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity
•170.074	11.241	• 78.802	11.254	•48.417	15.943	•36.597	11.562	•27.399	12.269	• 19.57	19.716
169.873	16.68	78.496	43.035	•48.395	15.402	36.441	21.861	26.567	24.926	•19.425	24.236
165.4	16.68	• 78.22	16.409	•45.606	23.761	• 36.421	15.422	• 26.527	17.828	19.175	44.534
144.297	21.291	• 78.164	11.25	• 45.473	20.872	• 34.534	20.526	• 26.413	18.147	16.755	22.141
134.498	13.672	• 77.207	35.25	45.427	17.597	• 34.464	12.279	•26.375	19.095	• 16.718	17.498
132.8	28.669	76.831	13.397	• 45.305	11.716	34.151	15.611	• 26.182	37.268	•16.627	13.462
131.062	10.445	• 76.766	11.49	44.745	40.605	• 34.064	12.247	• 26.151	27.552	16.579	11.348
130.243	28.504	• 76.435	15.679	• 42.62	13.804	•33.113	19.015	25.014	48.136	• 16.134	10.903
129.398	52.711	• 76.342	29.378	•42.579	19.946	• 31.919	11.658	24.922	25.414	• 16.021	11.285
129.376	38.624	• 76.275	24.408	• 42.471	18.412	•30.761	30.439	• 23.19	15.971	15.98	26.499
•128.903	65.828	• 76.238	21.083	42.432	30.206	29.635	58.48	23.052	30.928	• 15.881	23.456
•128.363	59.554	• 76.175	16.781	38.248	19.942	• 29.353	10.55	• 22.8	20.615	• 15.784	10.532
•128.122	12.041	67.629	44.703	• 38.21	22.589	29.174	20.324	• 22.777	17.709	• 15.343	10.807
•127.965	54.104	• 65.301	25.924	• 37.26	17.825	• 29.063	20.91	22.704	31.69	• 15.032	18.517
118.899	21.242	58.092	27.71	• 37.231	35.247	• 28.628	11.137	• 21.555	45.567	• 14.977	11.816
102.024	13.152	57.249	12.931	37.2	29.453	27.97	29.961	• 21.489	33.652	• 14.941	12.185
78.915	39.648	• 51.149	14.349	37.15	36.605	27.727	24.992	• 21.342	29.191		
		49.578	40.776	• 36.788	27.466	27.568	25.918	19.646	45.771		

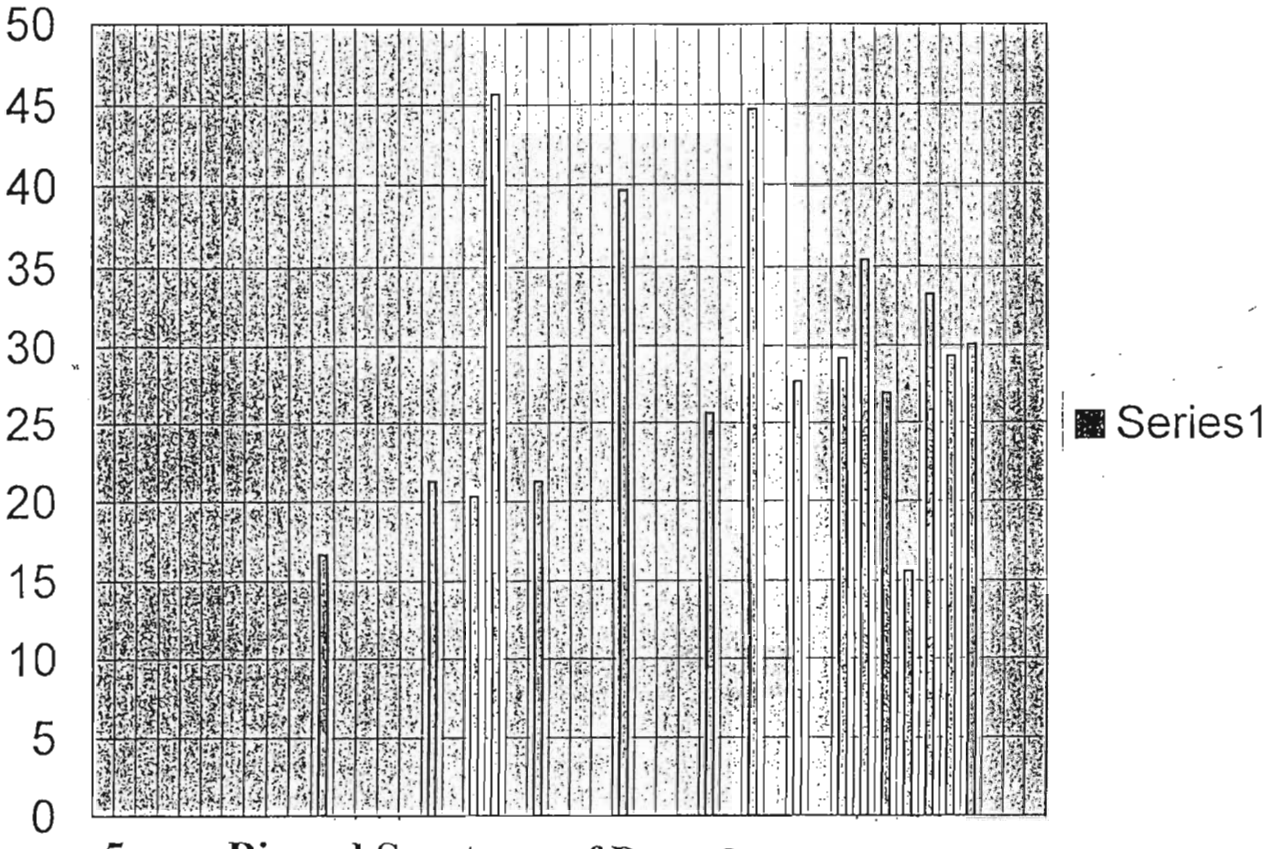
• - denotes extraneous peaks



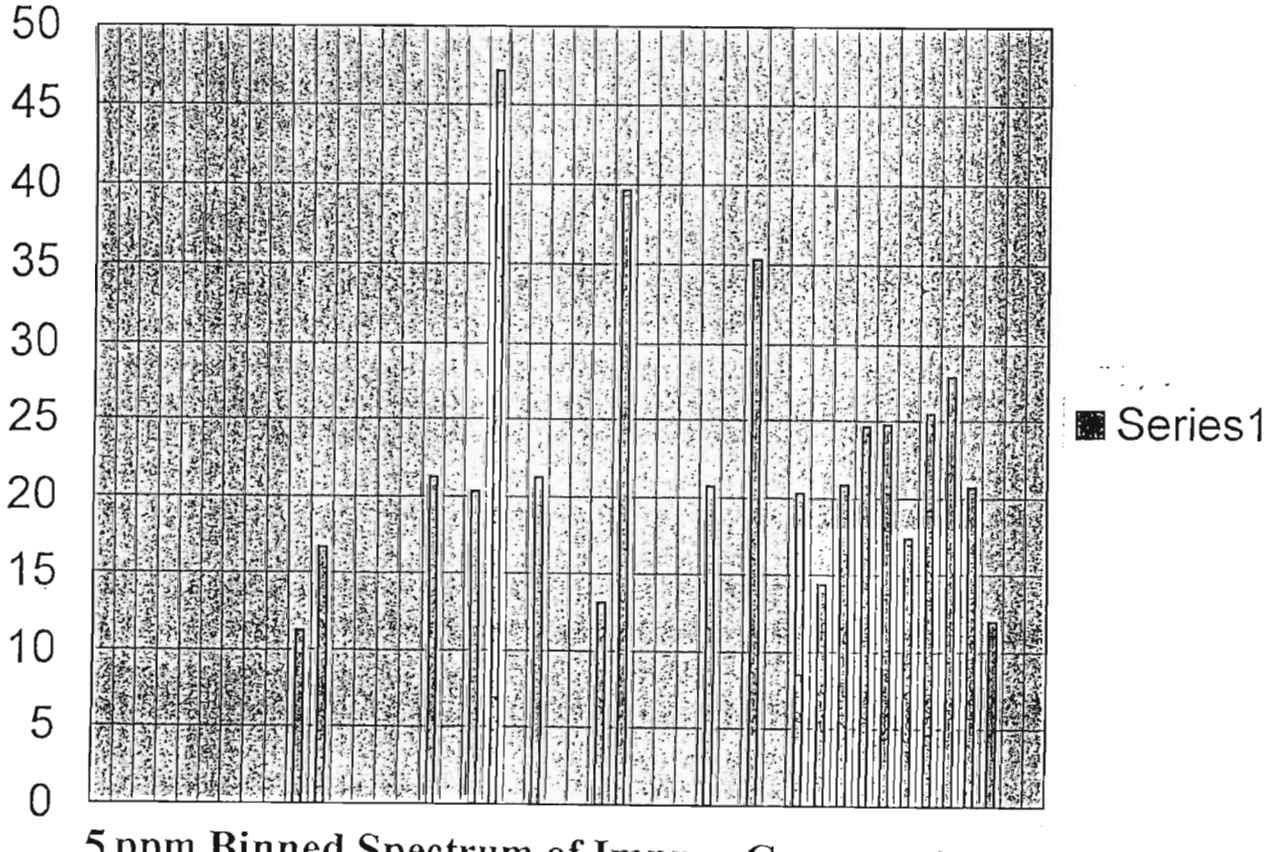
COMPOUND L



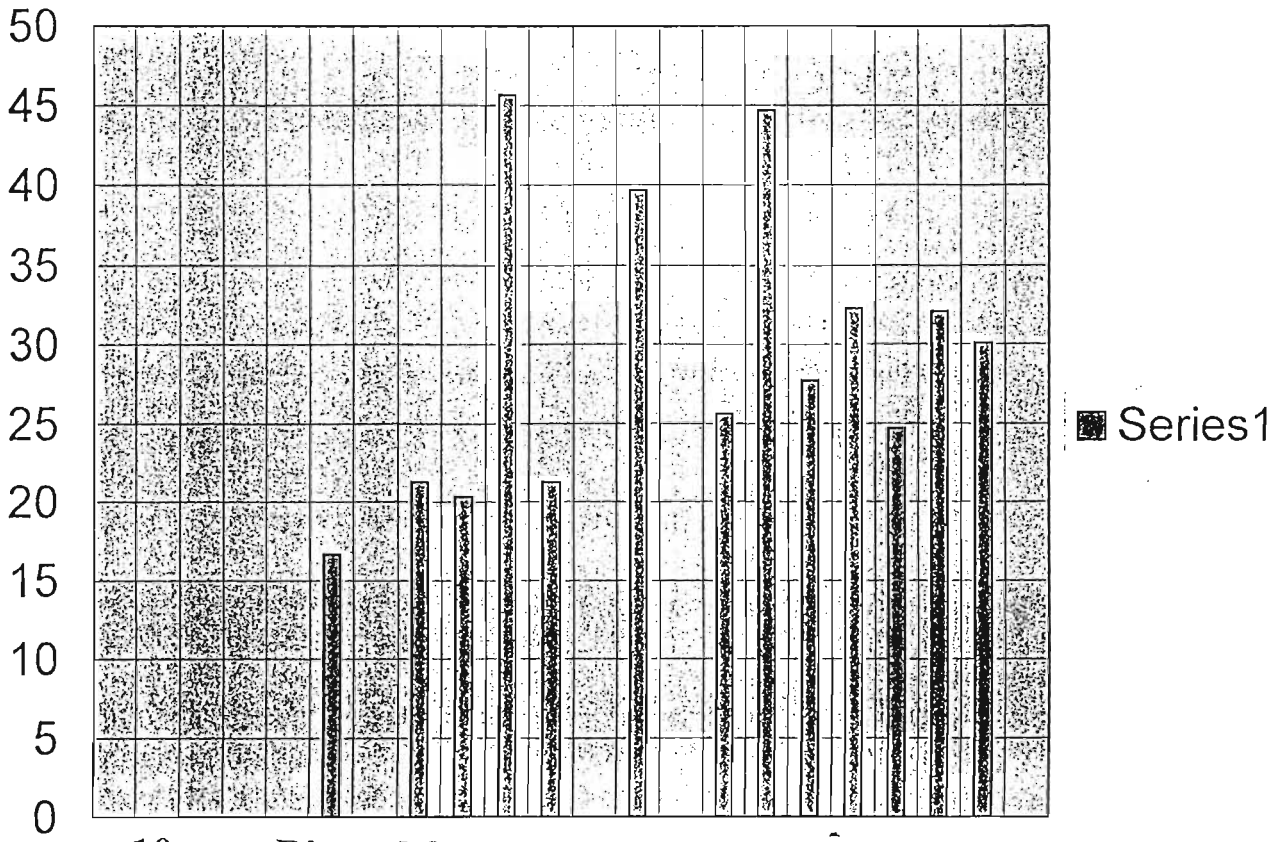
^{13}C NMR Spectrum of Compound L



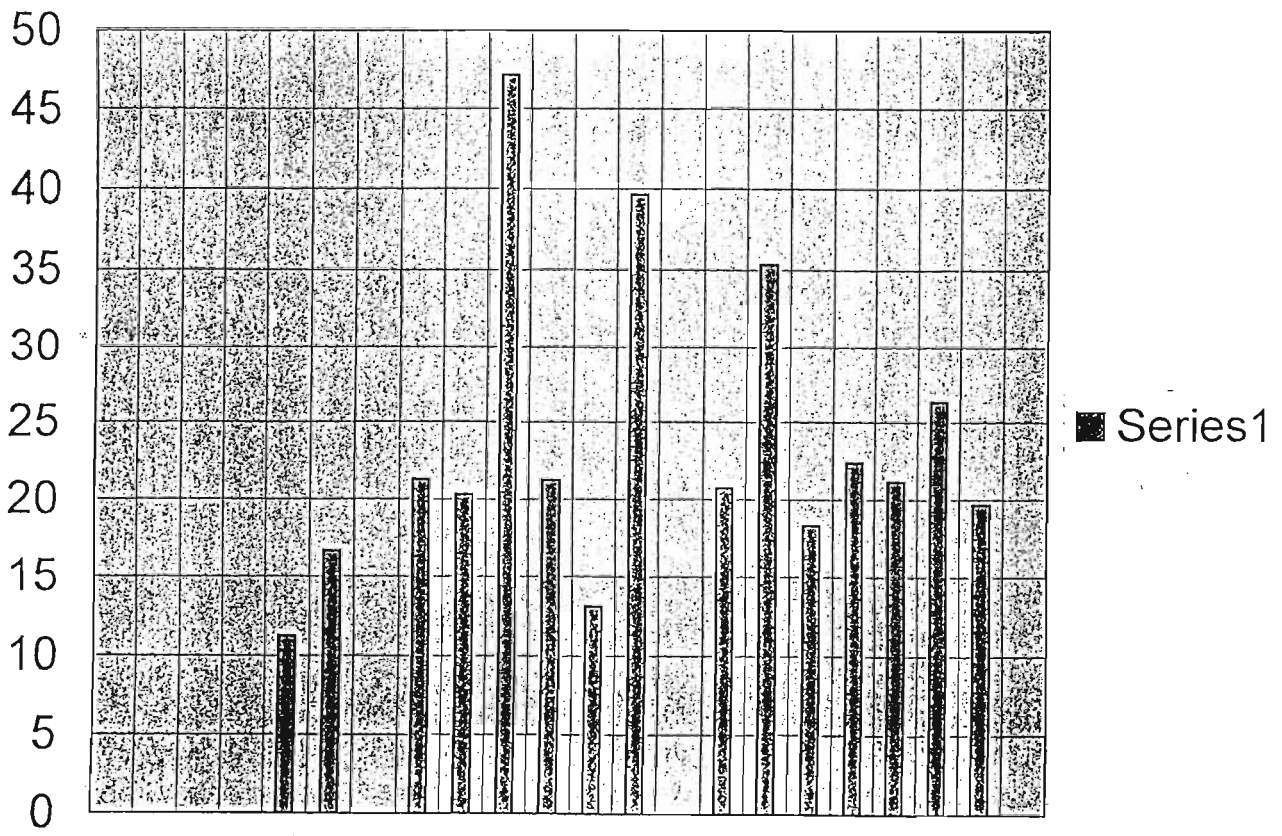
5 ppm Binned Spectrum of Pure Compound L



5 ppm Binned Spectrum of Impure Compound L



10 ppm Binned Spectrum of Pure Compound L

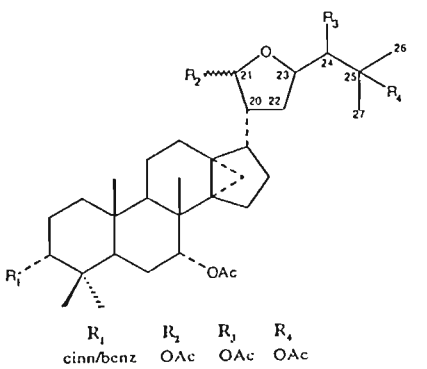


10 ppm Binned Spectrum of Impure Compound L

^{13}C NMR Data for Impure Compound LIII

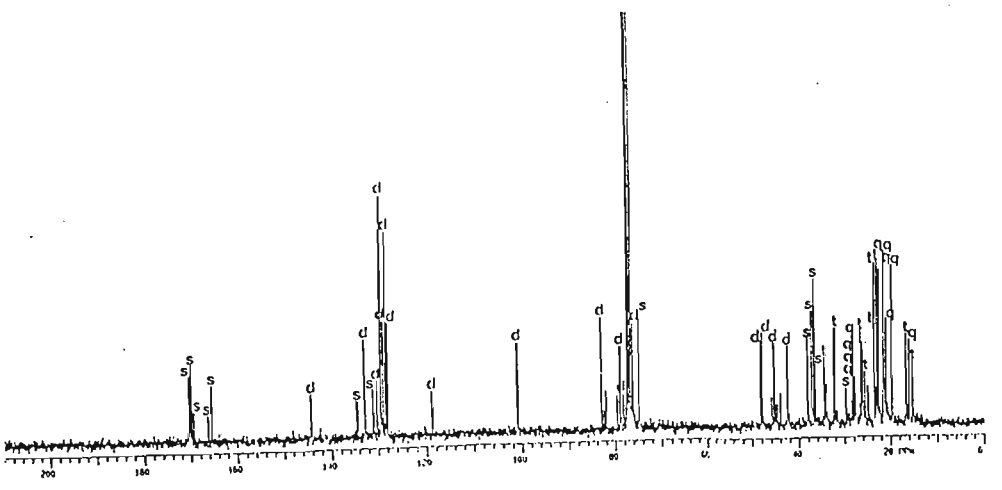
ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity
•170.535	11.372	128.356	78.808	48.184	33.224	• 36.402	23.071	• 25.563	21.372	• 19.618	61.997
170.424	28.052	• 128.111	10.276	47.832	37.122	34.464	25.98	• 24.892	15.532	16.645	36.407
170.31	24.327	127.945	44.088	• 45.895	11.071	• 34.086	11.105	23.46	63.271	• 16.499	11.513
170.103	33.012	118.854	17.433	45.349	32.564	• 34.001	16.511	• 23.138	15.129	• 16.014	13.252
• 169.976	12.783	100.956	35.347	• 45.251	18.777	• 32.646	11.463	22.98	35.729	15.897	34.415
169.791	28.961	• 96.242	12.782	• 43.958	12.37	32.104	37.938	22.777	63.347	• 15.829	19.59
• 169.333	11.044	82.955	45.163	• 42.539	19.795	29.672	13.86	• 22.66	37.531	15.217	24.49
165.565	20.461	• 82.065	14.47	42.409	33.263	• 28.346	16.131	22.38	60.704		
144.29	16.321	• 79.5	12.516	• 42.356	15.316	• 28.237	33.652	• 22.258	20.306		
134.479	13.505	78.862	31.592	• 38.137	18.047	27.985	33.57	21.52	34.034		
132.798	37.402	• 78.173	12.837	38.095	35.16	27.934	13.975	21.468	57.684		
131.035	19.996	77.198	24.797	• 37.886	13.666	27.705	17.797	• 21.402	67.201		
130.236	21.136	76.178	41.362	37.213	44.247	26.53	35.962	21.31	39.829		
129.378	92.42	• 76.115	23.778	37.167	21.866	26.136	32.249	• 21.163	17.945		
128.907	44.27	74.799	46.212	36.755	56.503	• 26.079	15.752	• 21.069	11.012		
128.413	30.055	• 48.247	17.987	• 36.574	36.609	• 25.98	26.353	20.844	46.375		

• - denotes extraneous peaks

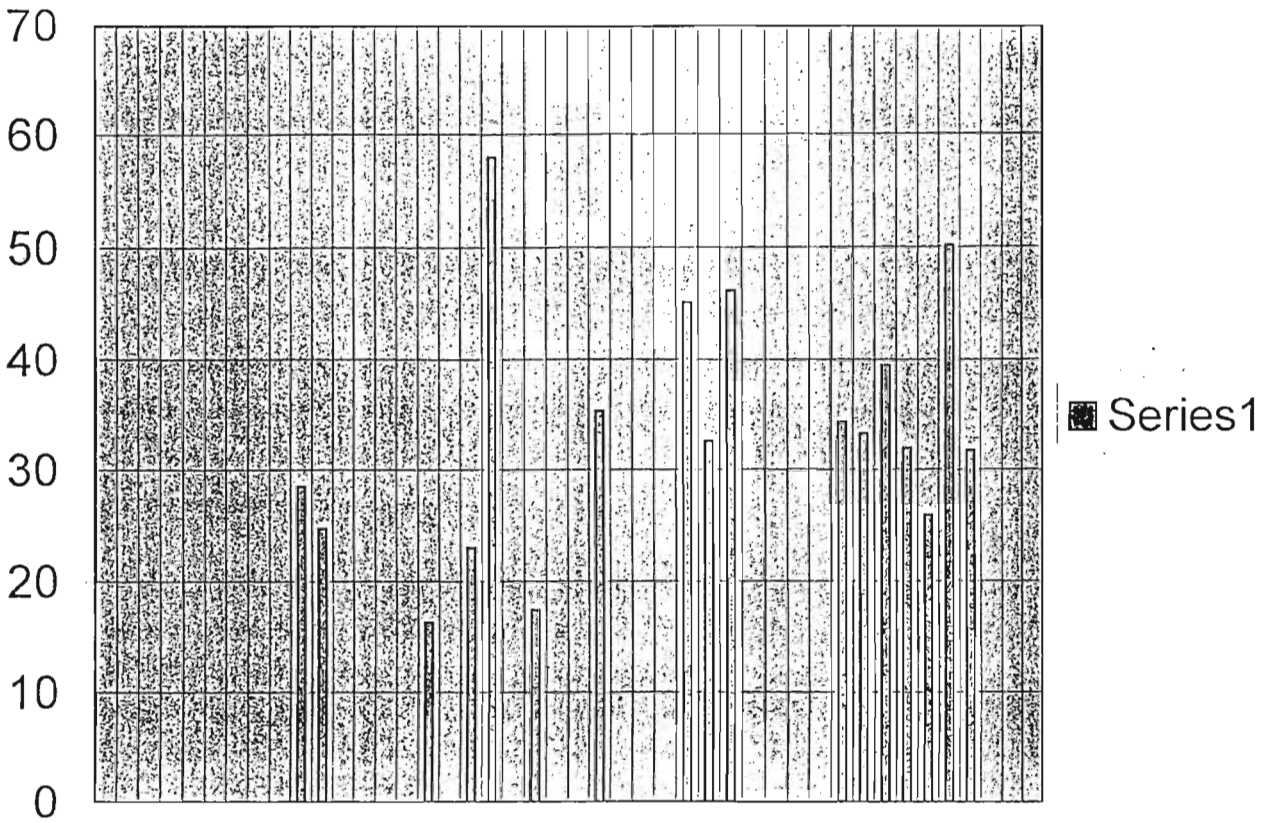


R_1	R_2	R_3	R_4
cinn/benz	OAc	OAc	OAc

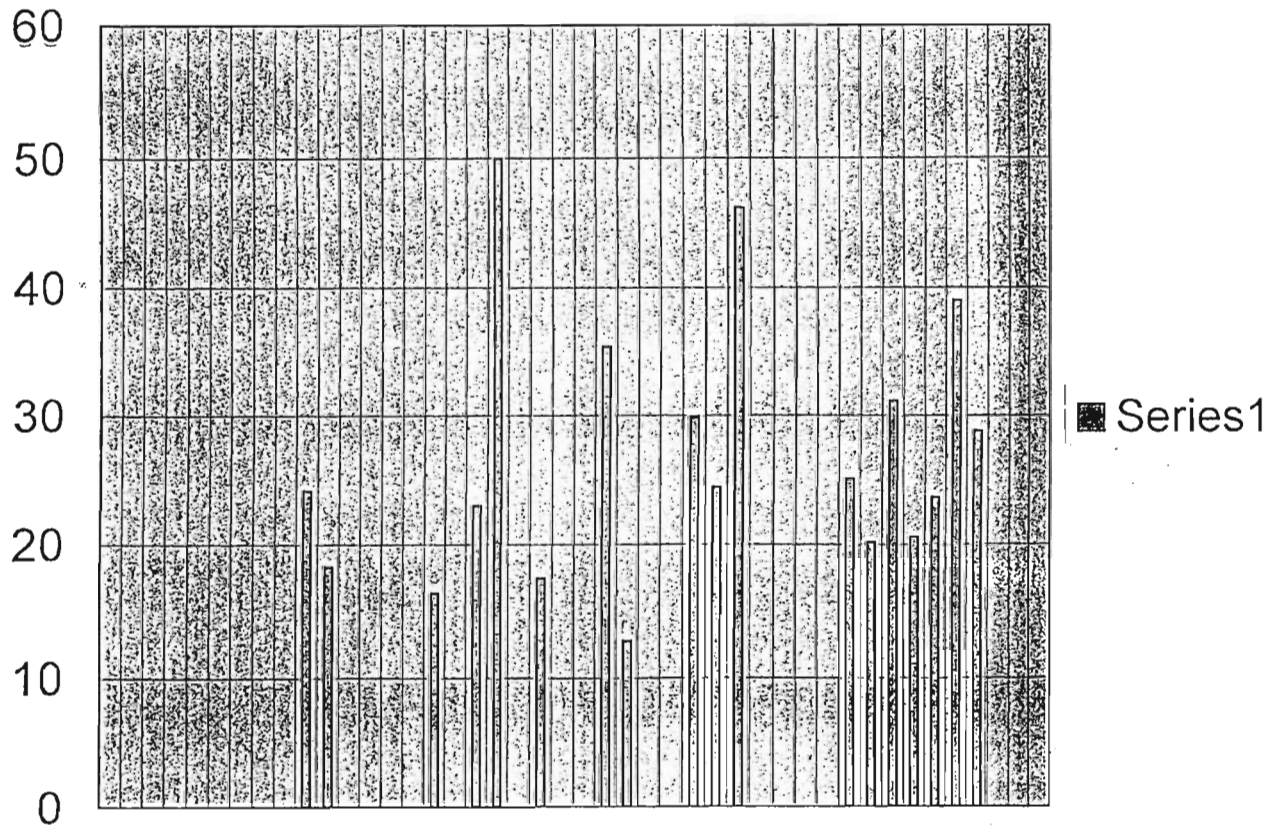
COMPOUND LIII



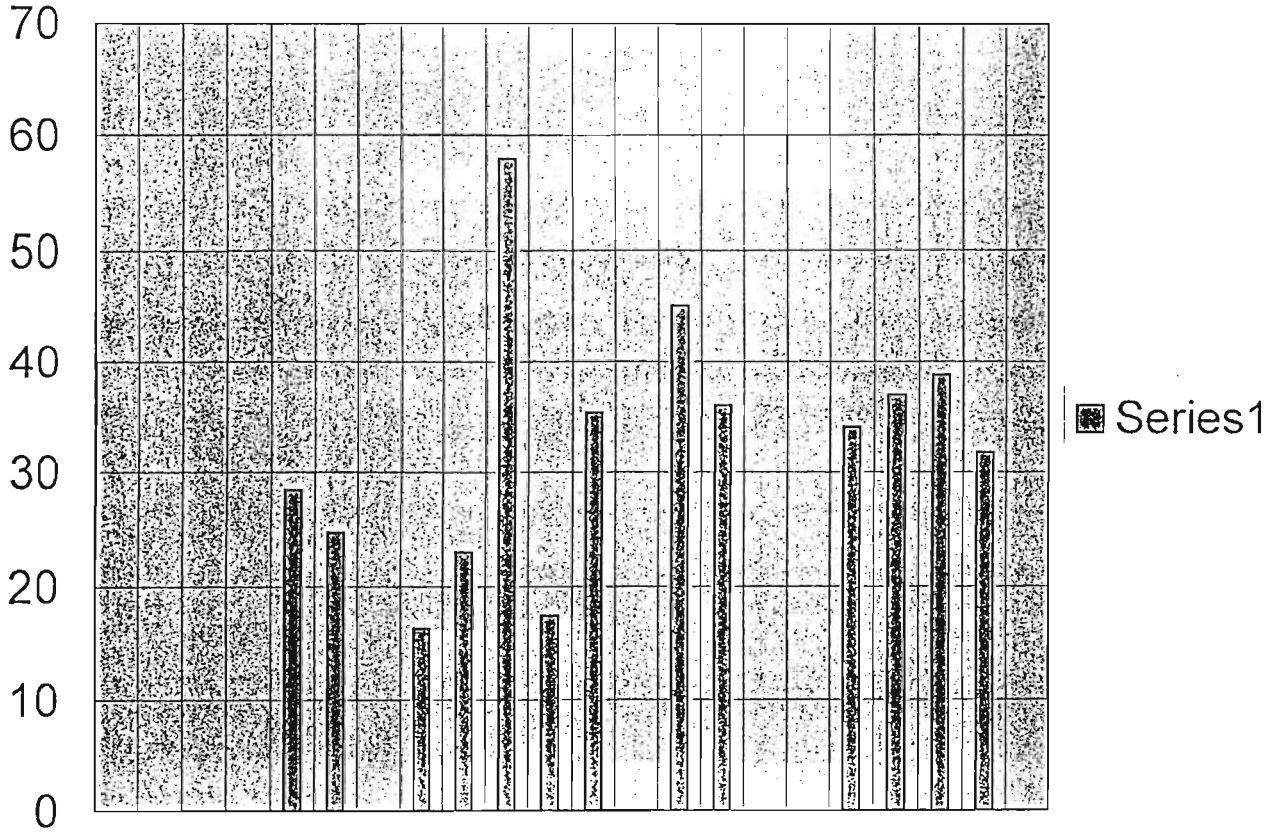
^{13}C NMR Spectrum of Compound LIII



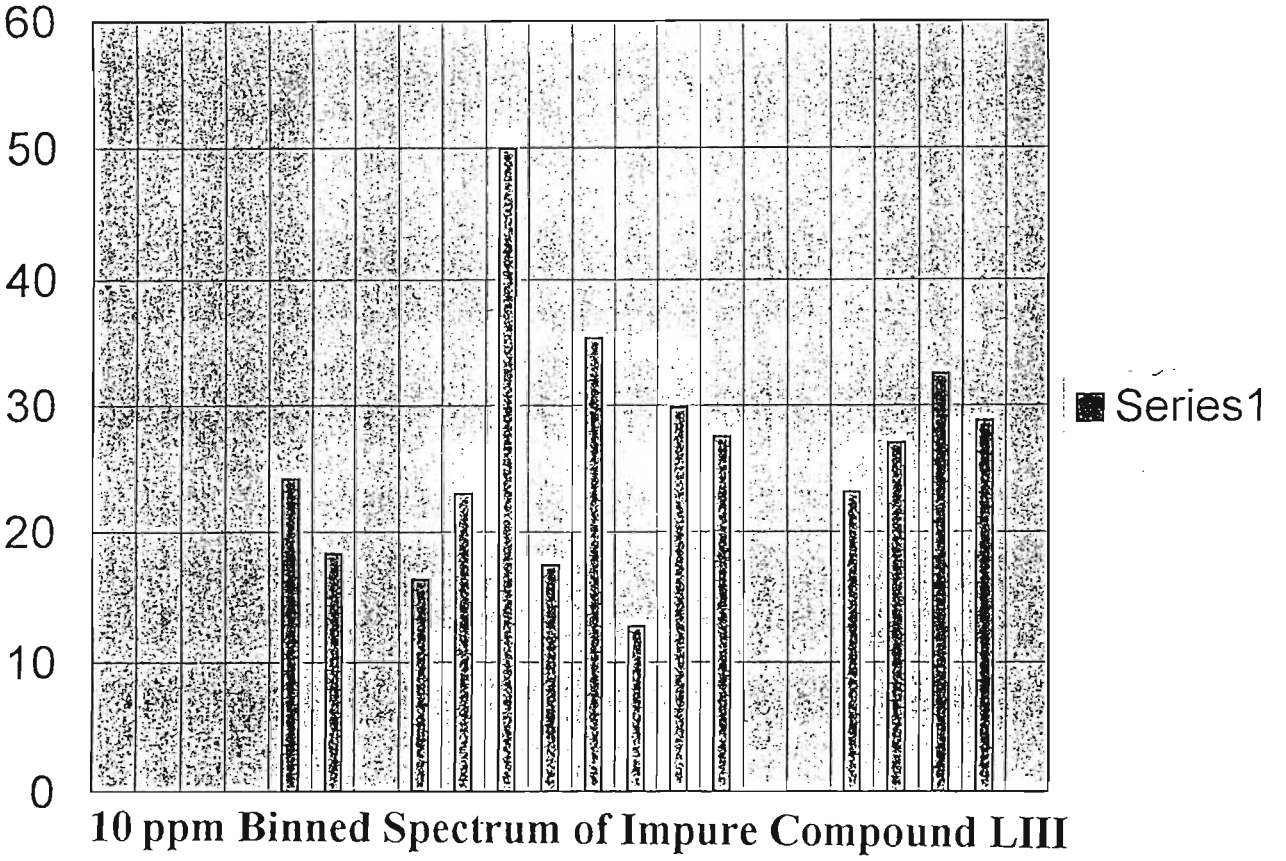
5 ppm Binned Spectrum of Pure Compound LIII



5 ppm Binned Spectrum of Impure Compound LIII



10 ppm Binned Spectrum of Pure Compound LIII



10 ppm Binned Spectrum of Impure Compound LIII

^{13}C NMR Data for Impure Compound XV

ppm	Intensity
171.545	70.936
• 124.815	11.328
• 112.77	9.703
• 112.643	11.006
102.361	18.178
• 98.414	30.038
• 97.638	18.229
• 80.373	12.41
• 78.895	20
78.608	38.426
78.321	80.395
• 77.926	36.71233
• 77.494	30.873
• 77.447	27.194
• 77.29	36.71233
• 76.651	36.71233
• 75.27	17.697
• 74.53	65.985

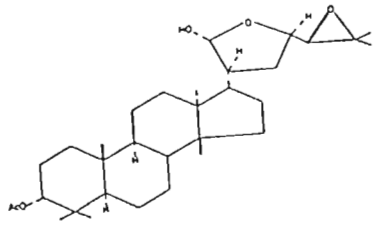
ppm	Intensity
• 74.406	30.867
• 73.812	18.974
67.878	36.452
• 65.482	16.845
58.356	30.441
• 57.56	16.082
• 50.825	18.218
• 49.616	13.511
49.442	32.545
• 48.815	18.903
• 48.366	17.257
• 44.981	29.182
44.836	36.172
44.086	53.731
• 43.924	26.53
41.344	79.147
• 41.253	12.32
• 39.136	25.386

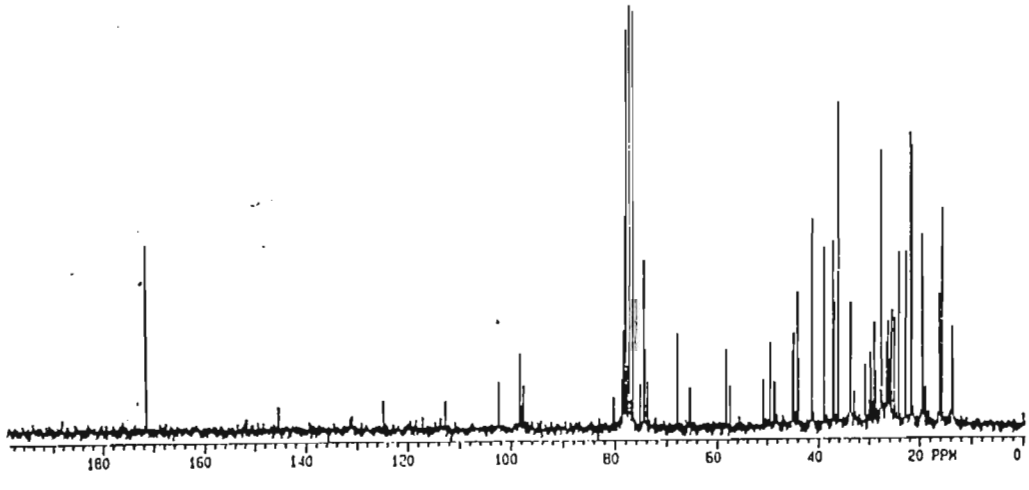
ppm	Intensity
39.063	74.159
37.371	38.529
37.312	79.982
• 37.064	35.053
37.007	50.488
36.245	126.424
33.837	50.152
• 33.727	22.038
• 33.074	14.666
30.881	23.971
• 30.076	10.152
• 29.763	29.578
• 29.657	15.681
29.043	40.217
• 28.976	32.514
• 28.679	21.565
27.696	104.58
• 27.549	36.955

ppm	Intensity
27.51	34.107
• 26.75	32.755
• 26.331	41.067
• 26.178	18.301
• 25.931	25.639
• 25.686	32.081
25.601	45.327
25.035	43.351
• 24.981	27.255
24.212	69.976
22.835	67.032
21.881	123.256
21.541	109.688
19.558	74.862
19.453	46.126
• 19.234	40.355
• 18.729	15.813
16.195	51.355

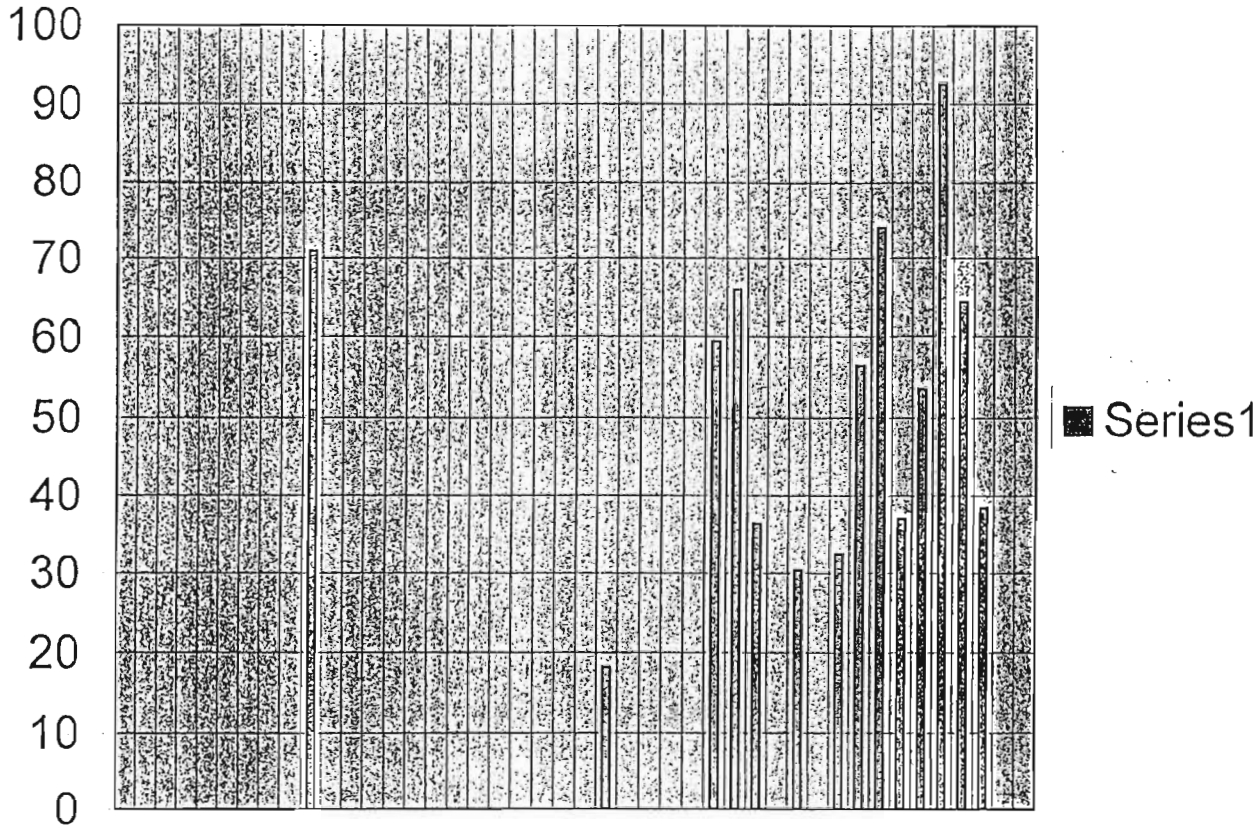
ppm	Intensity
• 16.05	20.673
15.7	85.256
• 15.591	27.056
13.739	38.489
• 13.541	12.828

• - denotes extraneous peaks

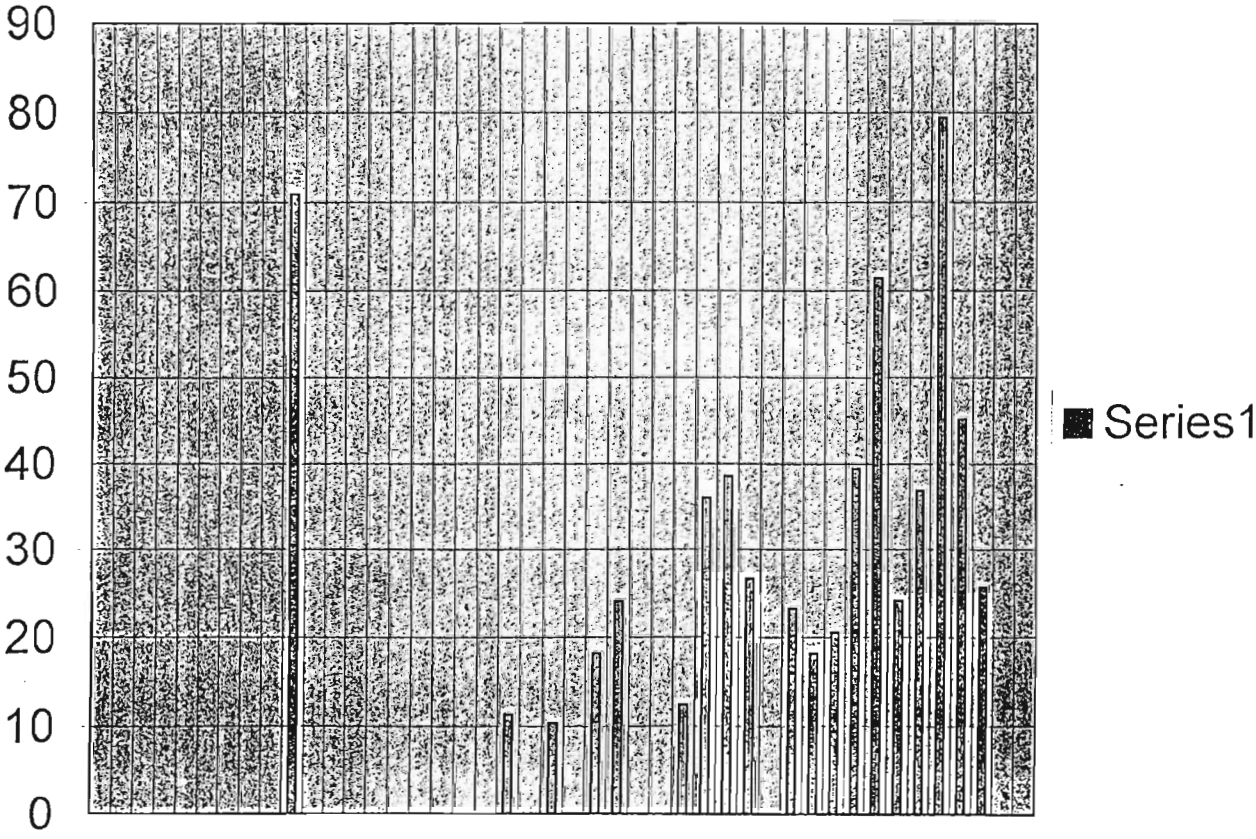




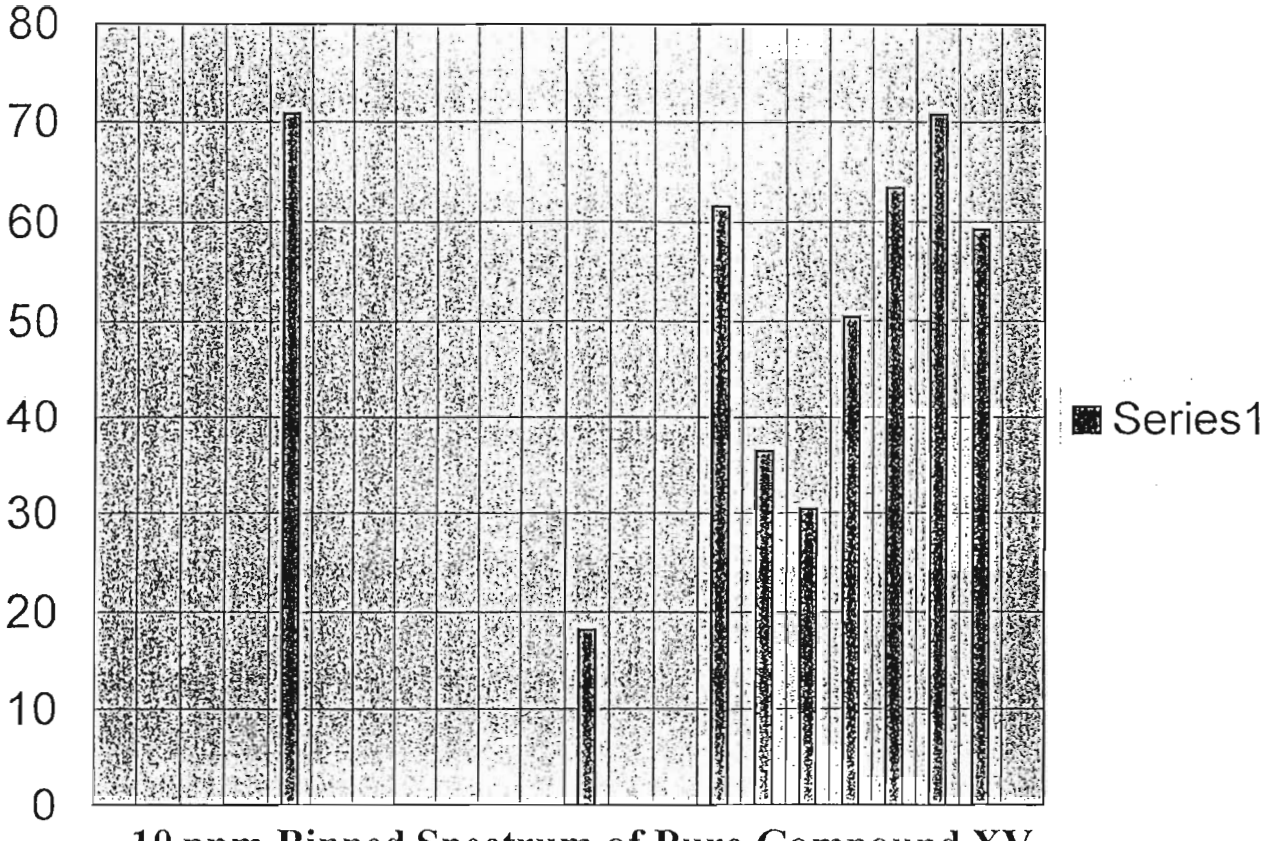
^{13}C NMR Spectrum of Compound XV



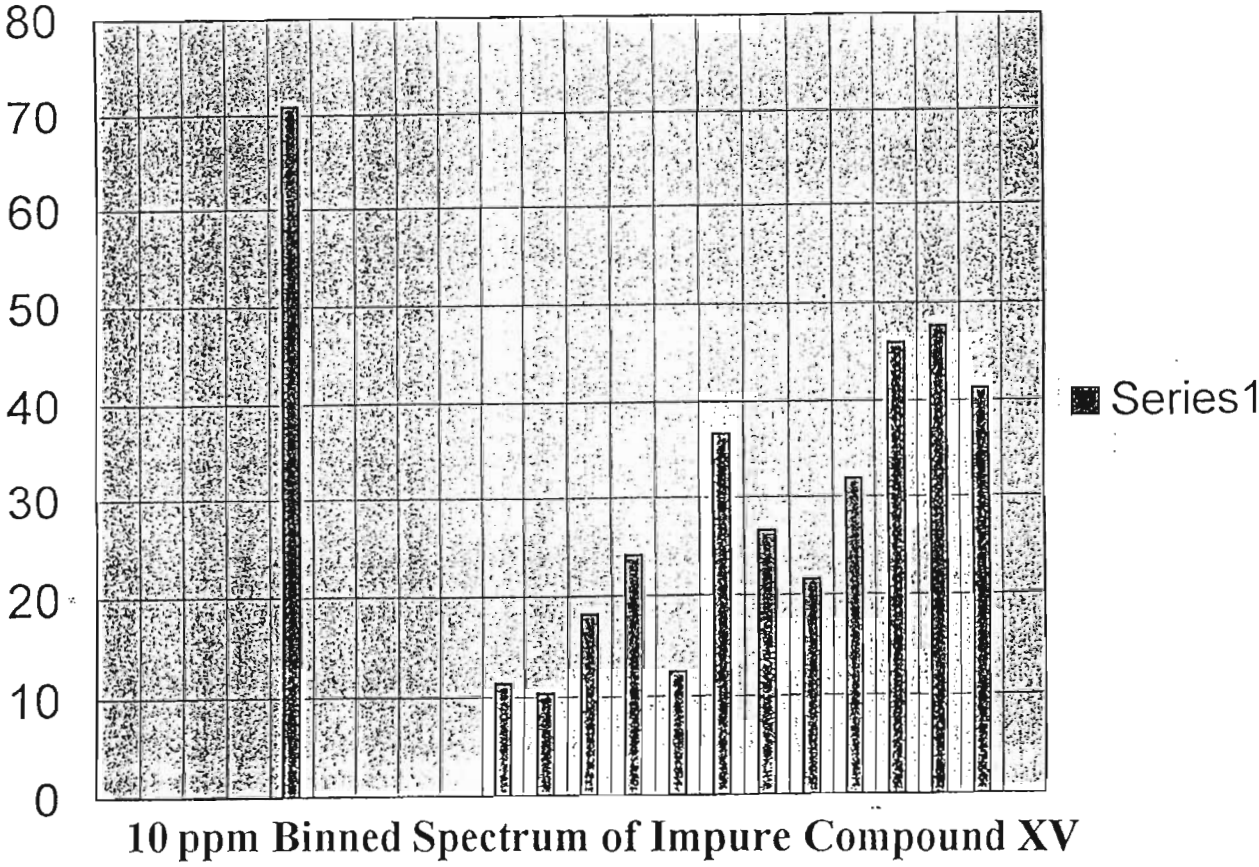
5 ppm Binned Spectrum of Pure Compound XV



5 ppm Binned Spectrum of Impure Compound XV



10 ppm Binned Spectrum of Pure Compound XV



10 ppm Binned Spectrum of Impure Compound XV

^{13}C NMR Data for Impure Compound XVI

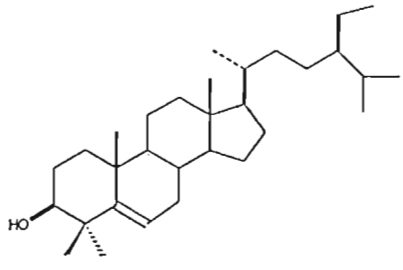
ppm	Intensity
141.156	27.412
•130.405	14.197
•130.086	10.172
122.085	46.53
•106.215	11.964
• 77.884	23.1645
• 77.734	15.15
• 77.684	10.232
• 77.447	21.918
• 77.245	23.1645
• 76.608	23.1645
72.018	45.358
• 62.289	12.449
56.911	40.723
•56.233	11.849
56.185	13.812
• 56.124	18.892

ppm	Intensity
50.242	41.783
• 45.932	13.724
42.41	50.821
39.872	39.097
37.345	44.503
36.599	30.116
36.249	25.572
• 36.03	9.533
• 35.865	17.105
• 34.757	13.459
• 34.312	11.966
• 34.193	11.442
• 34.144	11.56
• 34.075	10.396
34.019	16.896
• 33.887	12.21
• 33.792	13.429

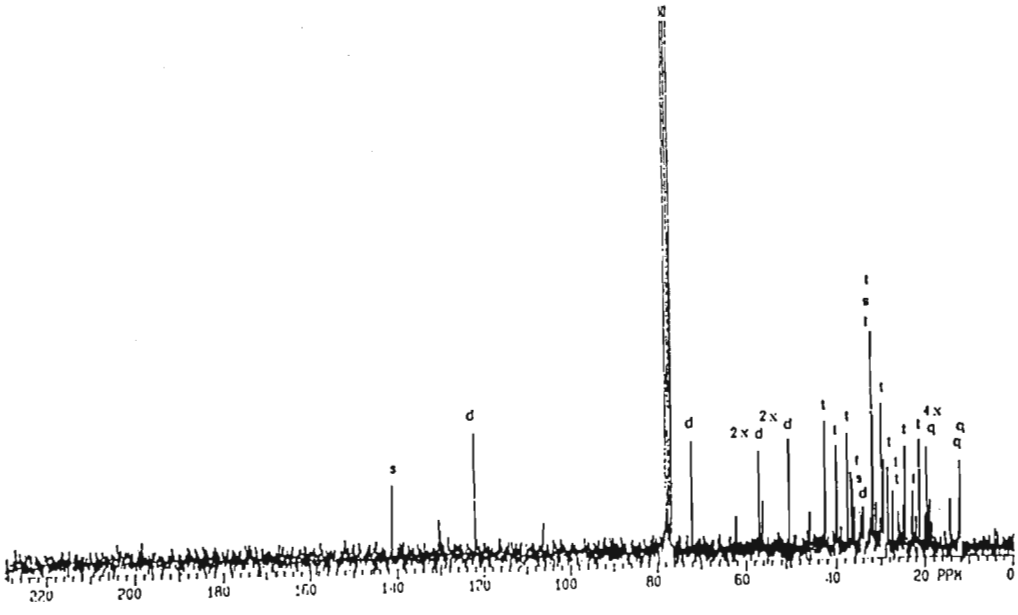
ppm	Intensity
• 32.516	11.477
32.136	10.803
31.983	91.911
31.738	50.479
• 31.044	17.181
• 30.333	11.097
• 29.793	56.21
• 29.625	18.58
• 29.421	28.324
29.188	33.293
28.333	30.59
• 27.294	21.525
26.087	14.098
• 25.988	9.605
• 24.914	15.893
24.366	37.966
23.106	11.135

ppm	Intensity
• 22.765	20.577
• 22.069	11.157
• 21.932	11.711
21.135	45.312
• 20.274	11.042
19.892	12.748
19.463	40.577
19.077	15.819
18.836	16.574
• 18.757	17.75
• 14.182	18.545
• 12.012	14.733
11.897	33.488

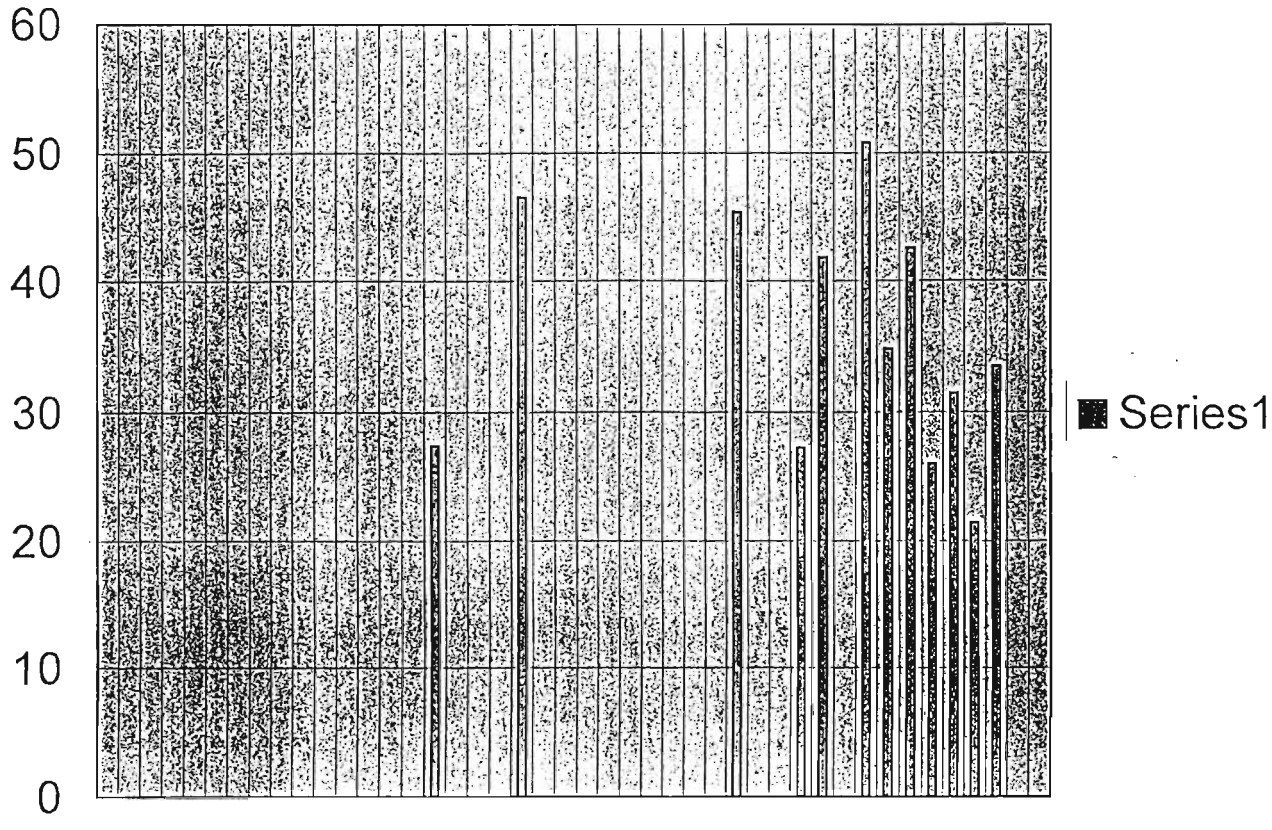
• - denotes extraneous peaks



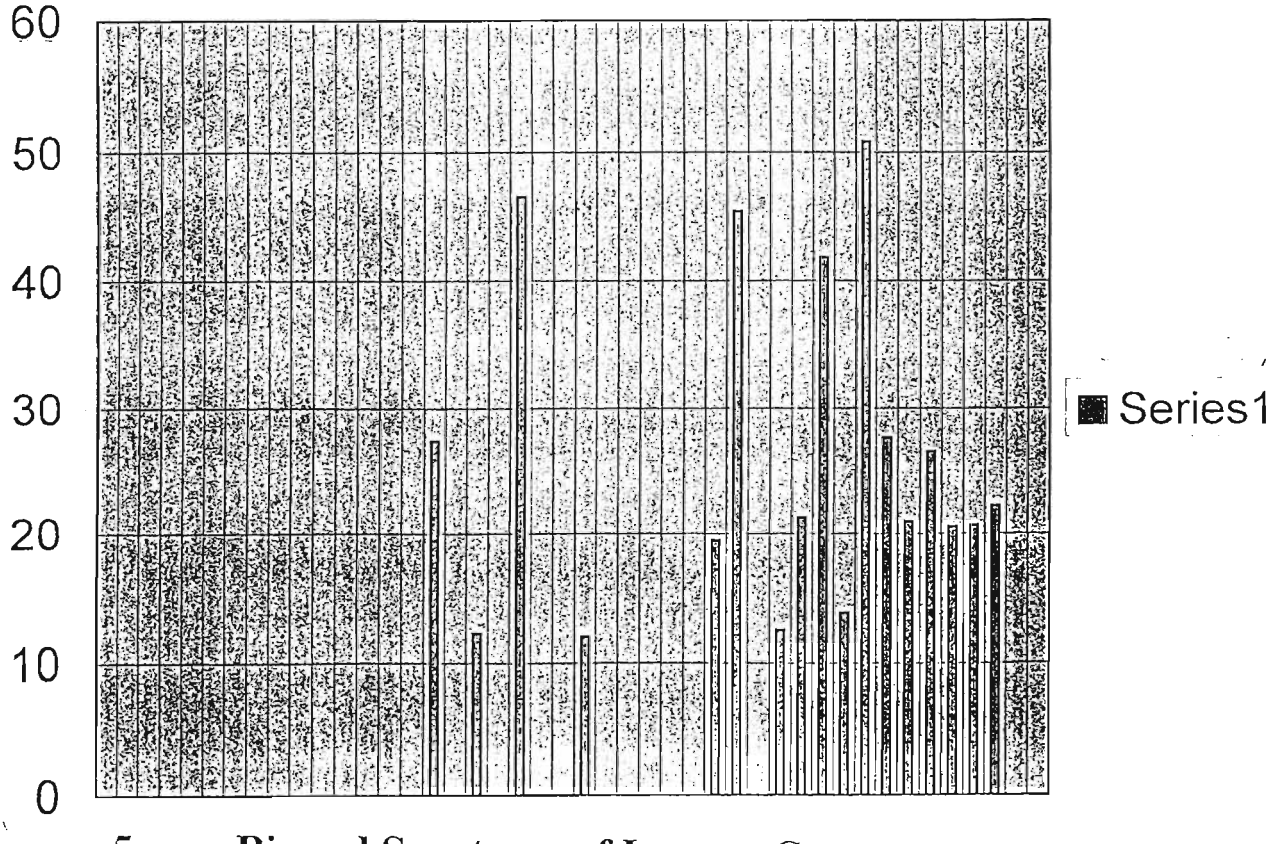
SITOSTEROL (XVI)



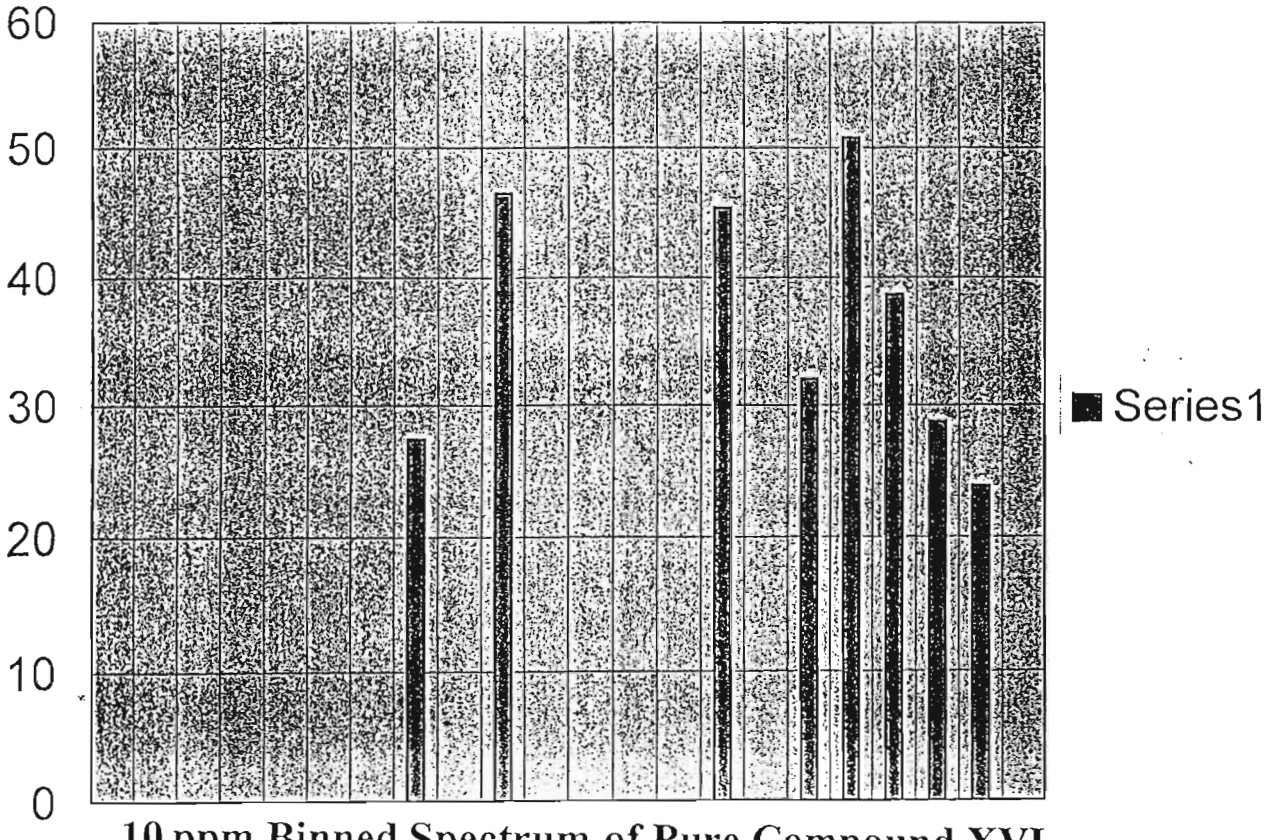
^{13}C NMR Spectrum of Compound XVI



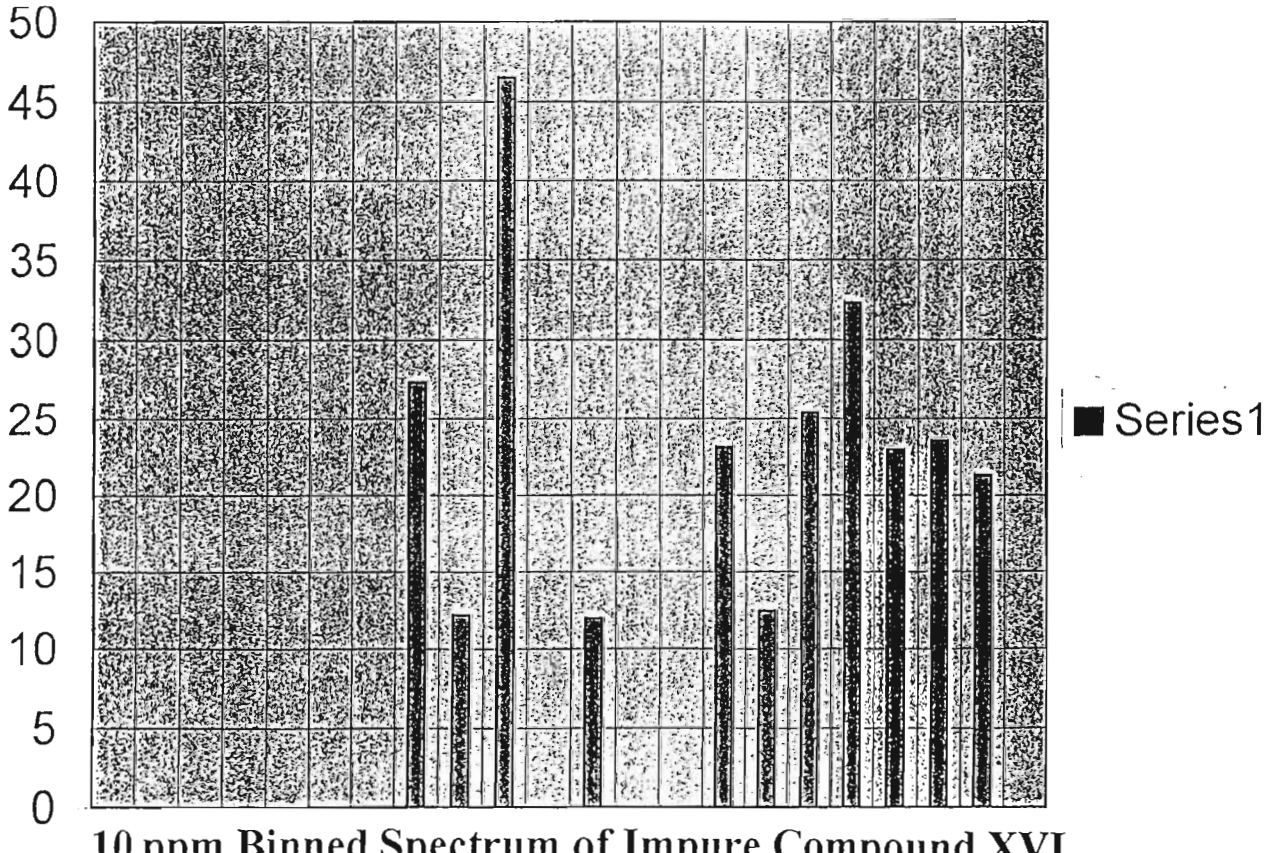
5 ppm Binned Spectrum of Pure Compound XVI



5 ppm Binned Spectrum of Impure Compound XVI



10 ppm Binned Spectrum of Pure Compound XVI



10 ppm Binned Spectrum of Impure Compound XVI

^{13}C NMR Data for Impure Compound XXXXIV

ppm	Intensity
217.538	50.395
173.361	51.683
• 173.317	20.45
• 170.327	14.28
170.218	60.156
• 159.193	13.698
147.249	75.192
• 147.118	17.104
125.892	74.614
• 118.909	15.242
75.554	90.784
• 75.253	18.849
• 60.442	19.38
52.233	92.303
• 48.185	20.051
46.966	97.686
• 46.819	15.019

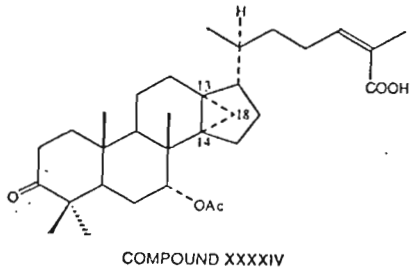
ppm	Intensity
46.733	110.248
• 46.527	12.244
44.618	102.268
• 42.917	17.154
• 41.906	16.574
39.55	72.269
• 38.802	14.029
37.724	93.274
• 36.948	11.917
36.905	23.596
36.576	99.783
• 36.395	12.243
35.517	98.615
• 35.183	20.362
35.058	104.224
• 34.663	17.112
• 34.355	11.692

ppm	Intensity
• 34.108	12.794
33.909	91.212
• 33.627	21.729
33.047	72.432
• 29.689	30.936
27.834	75.358
27.612	76.783
• 27.529	97.935
• 27.096	14.673
26.901	25.912
• 26.831	13.299
• 26.601	19.824
• 26.422	73.07
26.29	109.267
• 25.856	23.715
• 25.341	15.728
24.98	65.008

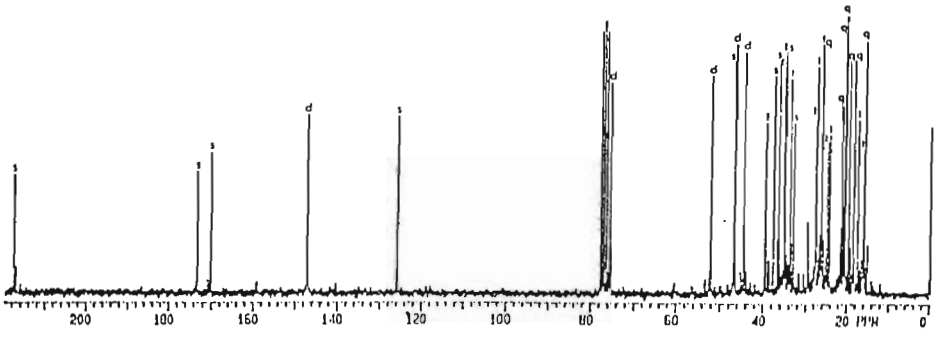
ppm	Intensity
24.404	67.436
• 24.29	19.641
• 21.927	17.358
• 21.854	12.65
21.527	79.567
• 21.256	18.823
• 21.008	36.781
20.828	110.904
20.563	118.541
19.784	102.404
• 19.257	20.425
18.604	99.844
• 18.522	28.534
• 18.358	14.498
17.53	74.037
• 16.796	14.597
16.231	61.324

ppm	Intensity
15.931	107.125
• 15.62	11.516
• 15.16	21.332
• -0.004	75.141

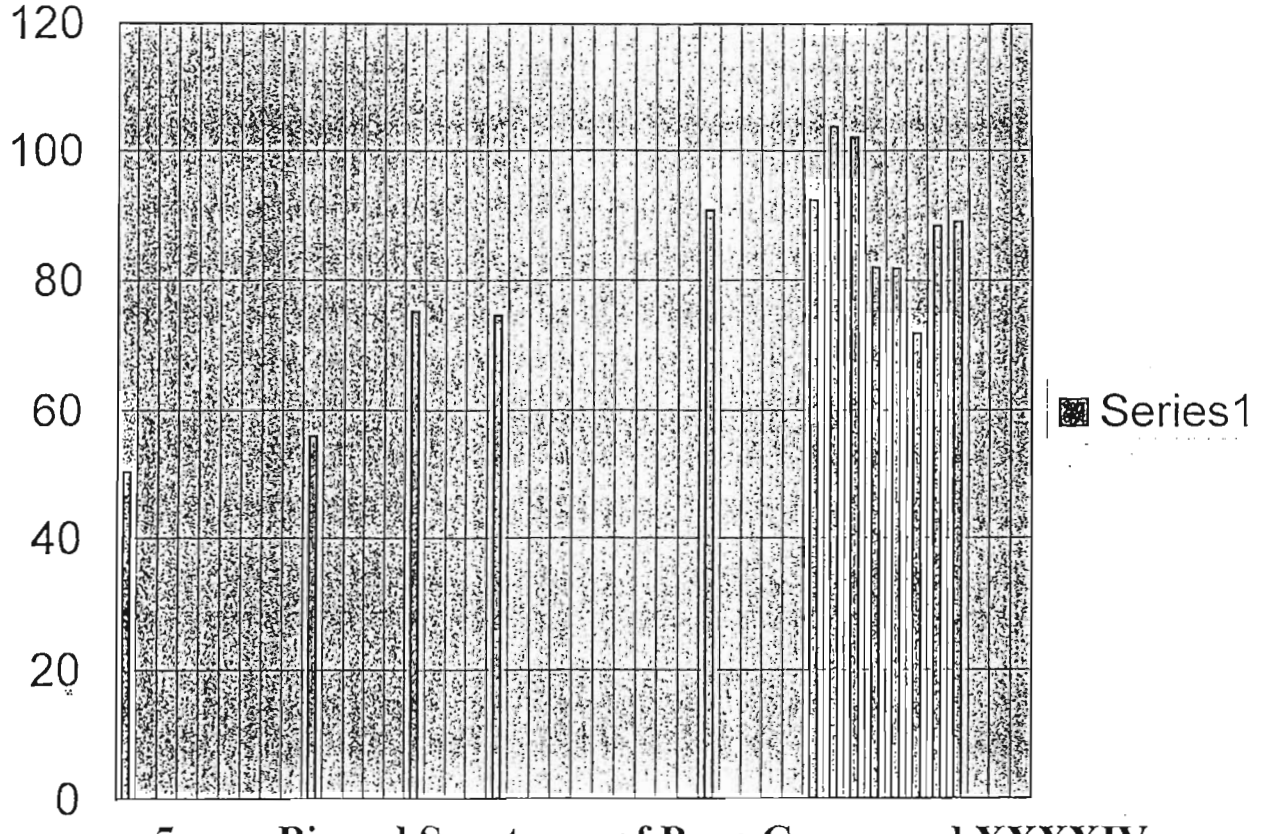
• - denotes extraneous peaks



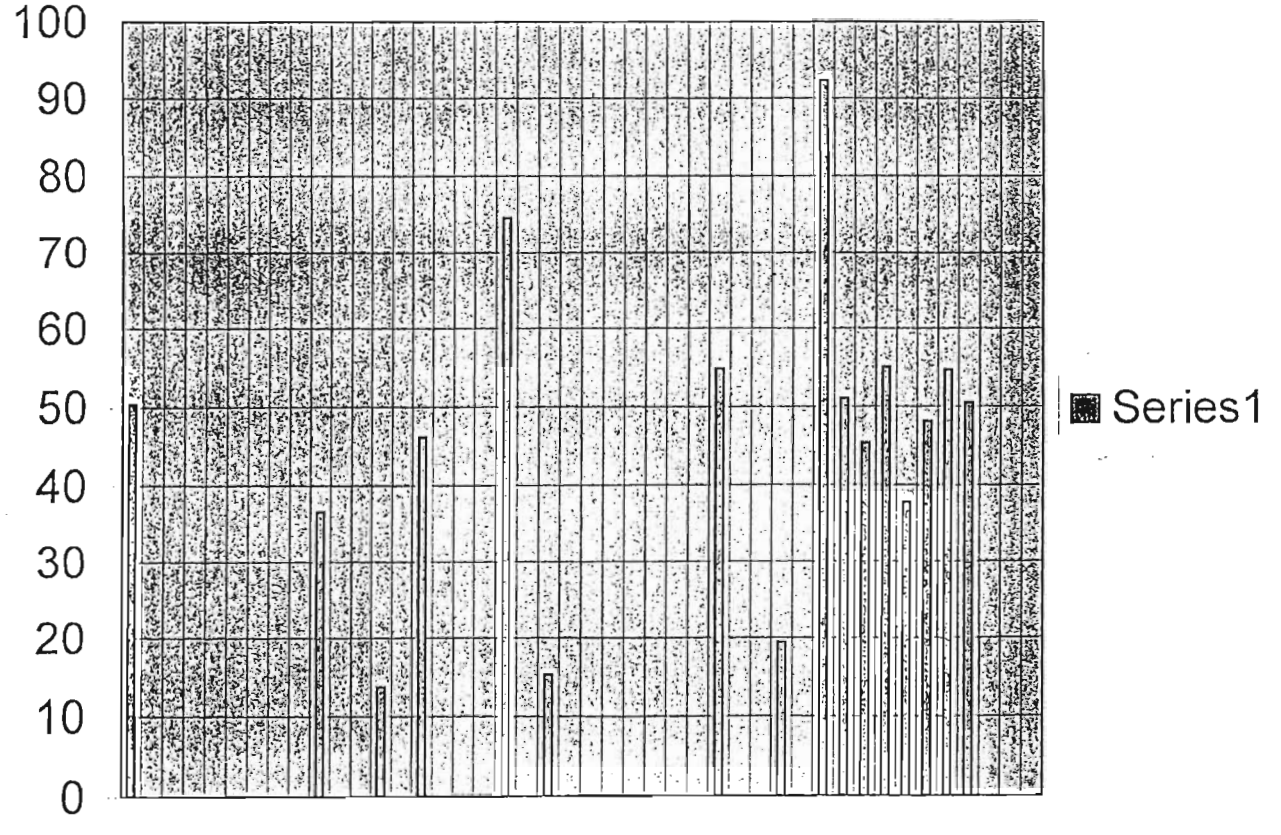
COMPOUND XXXXIV



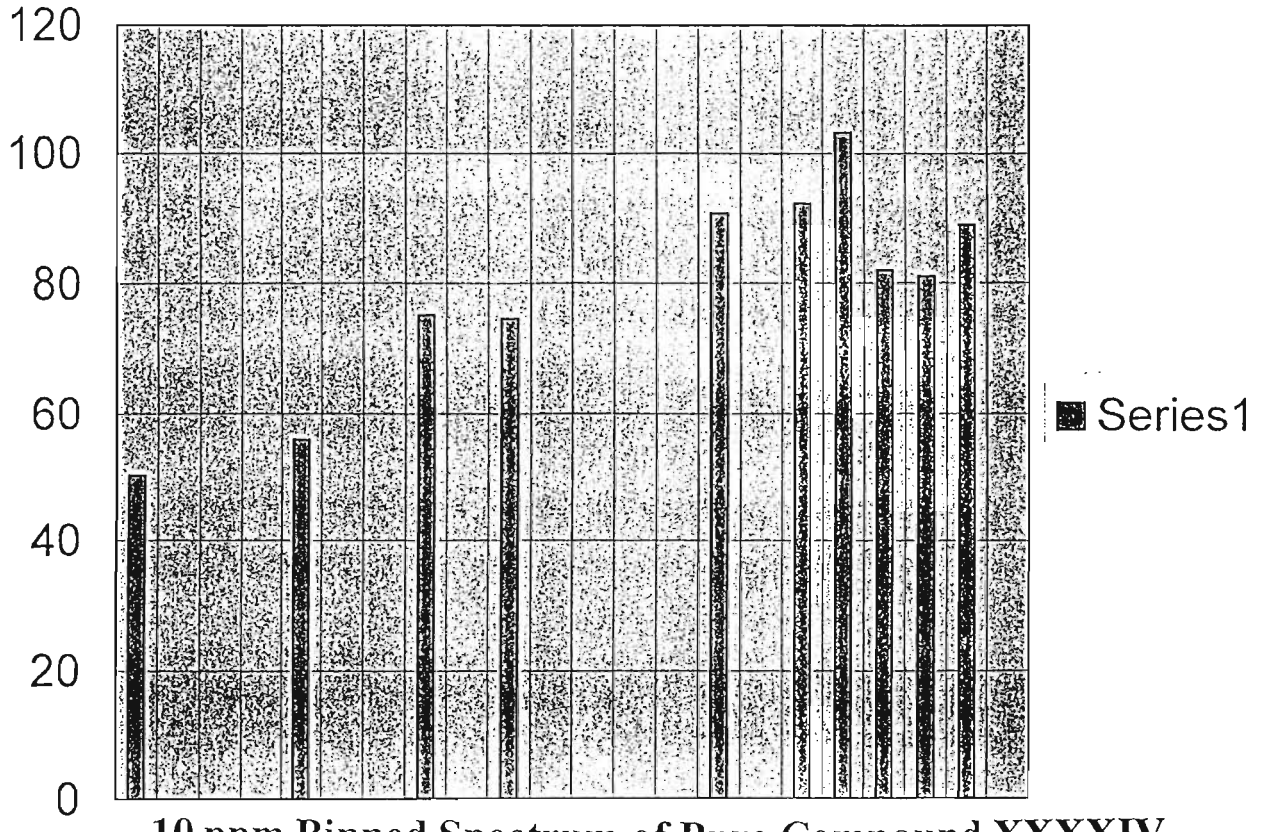
^{13}C NMR Spectrum of Compound XXXXIV



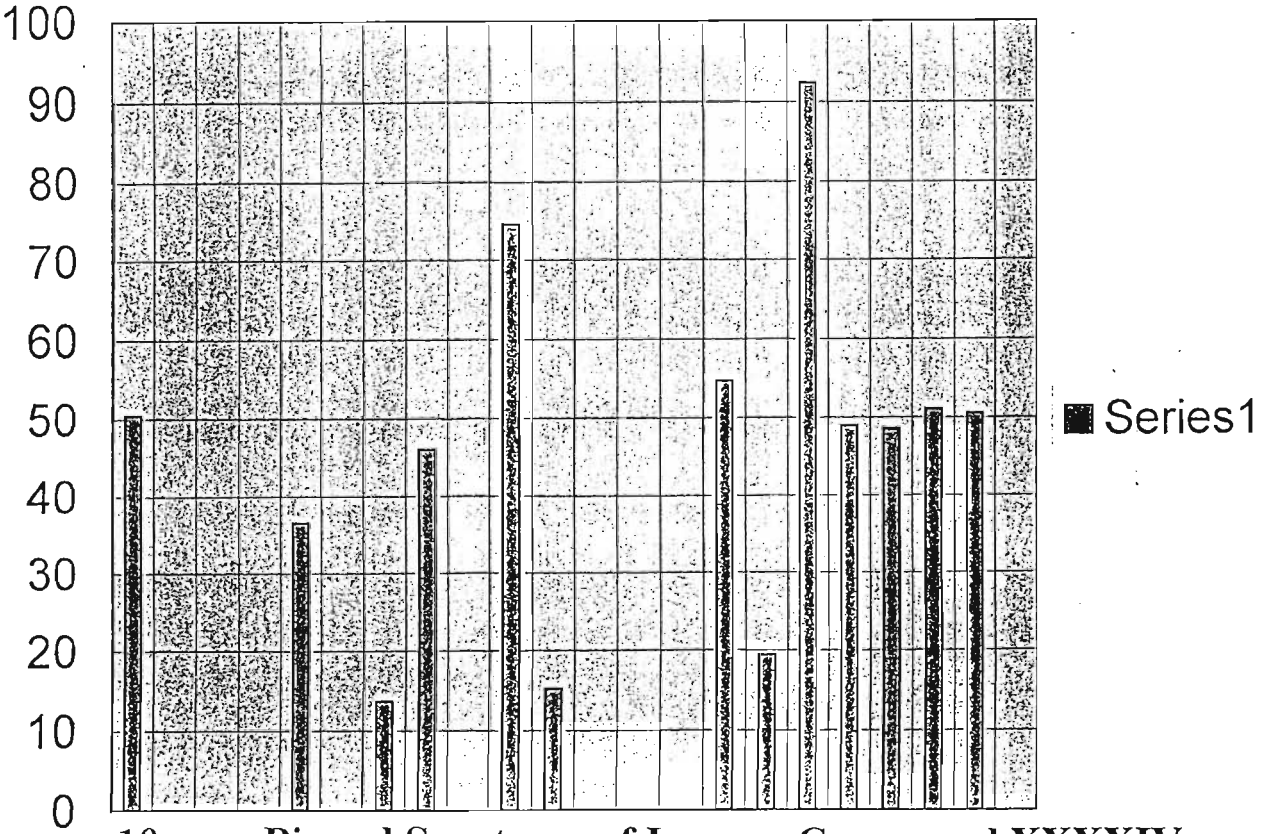
5 ppm Binned Spectrum of Pure Compound XXXXIV



5 ppm Binned Spectrum of Impure Compound XXXXIV



10 ppm Binned Spectrum of Pure Compound XXXXIV



10 ppm Binned Spectrum of Impure Compound XXXXIV

^{13}C NMR Data for Impure Compound LII

ppm	Intensity
170.983	31.578
170.127	21.654
• 169.978	16.801
169.791	30.506
166.372	10.797
• 166.337	12.781
165.564	17.548
• 144.338	14.017
144.296	16.552
134.476	14.506
• 134.445	10.326
132.798	36.687
131.026	20.737
130.238	23.604
129.379	86.643
128.91	59.983
128.366	67.321
127.955	54.192

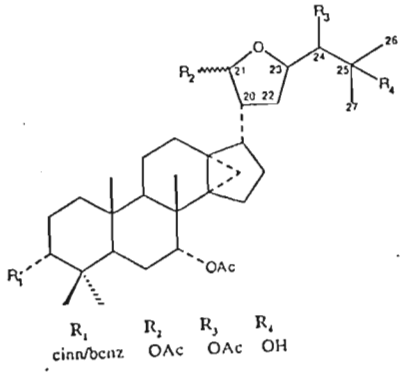
ppm	Intensity
118.846	18.567
100.639	36.508
• 96.292	17.982
82.559	18.146
78.845	29.185
78.162	16.381
• 78.073	12.246
• 77.271	12.851
• 77.202	23.912
• 77.149	12.939
76.33	48.298
• 76.257	10.848
• 76.162	41.589
• 76.093	28.254
72.386	44.929
• 71.419	17.877
• 66.222	13.639
• 48.055	15.425

ppm	Intensity
47.981	30.074
47.774	35.109
• 45.756	11.091
45.33	29.901
• 45.232	13.9
• 43.906	16.006
• 42.524	19.215
• 42.493	17.074
42.4	36.125
• 38.123	18.662
38.079	33.641
37.199	53.037
37.16	23.95
36.757	51.83
• 36.648	31.542
• 36.402	30.525
34.472	24.639
• 34.177	10.981

ppm	Intensity
• 34.098	14.762
• 33.964	18.787
• 32.661	16.842
31.566	27.813
29.675	14.621
• 28.281	16.897
28.166	29.544
27.986	31.955
• 27.701	21.908
27.308	51.745
• 26.796	16.577
26.721	49.864
• 26.621	33.88
• 26.282	32.088
26.148	40.113
• 25.923	21.611
• 25.666	14.72
• 25.568	22.351

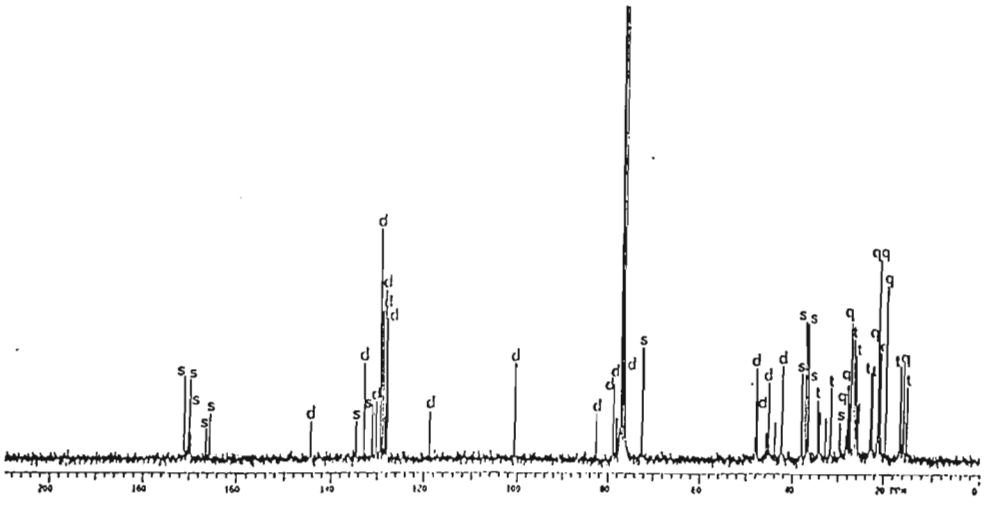
ppm	Intensity
23.142	20.319
22.989	34.126
• 22.76	18.819
22.663	33.649
21.537	42.021
21.475	46.853
21.336	79.417
• 21.173	23.627
20.862	44.307
19.635	67.845
16.654	36.296
• 16.489	12.499
15.909	37.28
• 15.845	21.931
15.248	28.582

• - denotes extraneous peaks

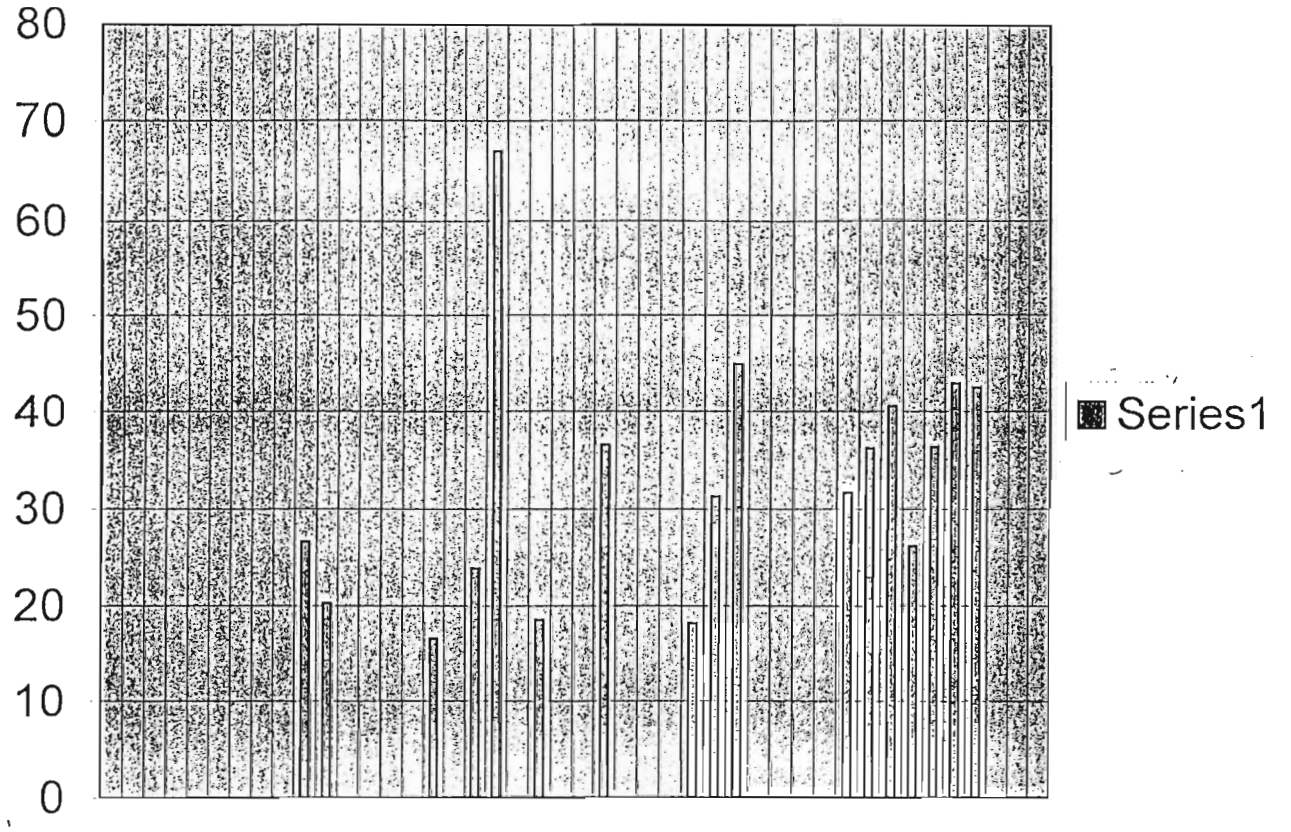


R_1	R_2	R_3	R_4
cinn/bcnz	OAc	OAc	OH

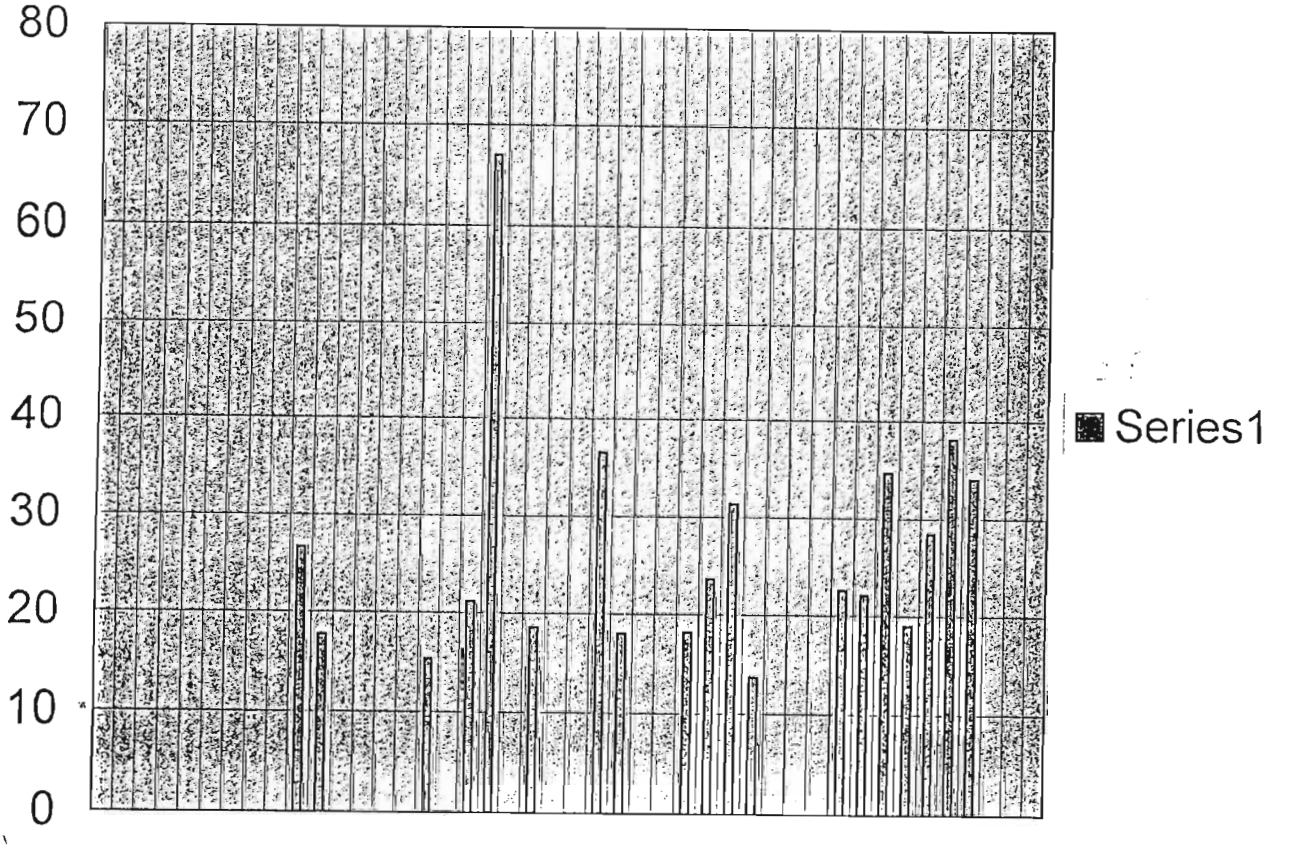
COMPOUND LII



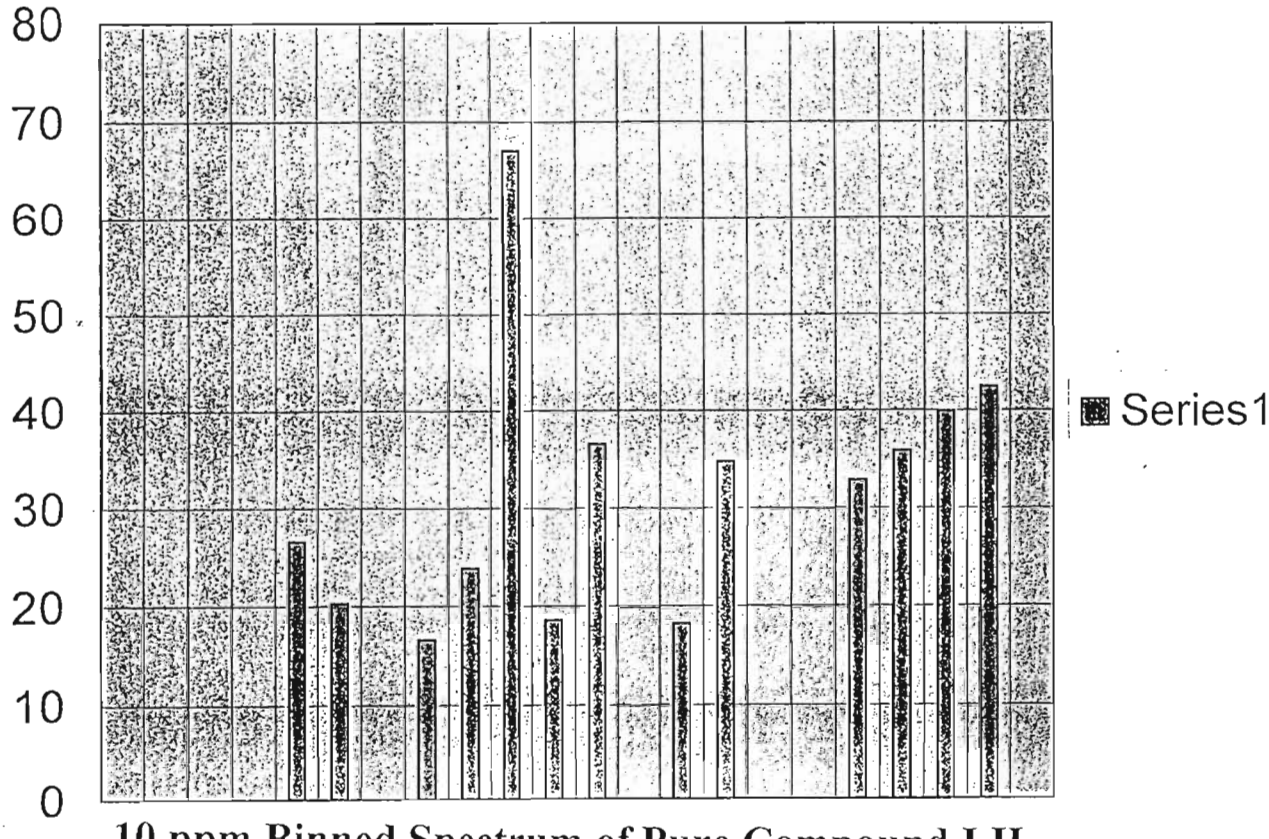
^{13}C NMR Spectrum of Compound LII



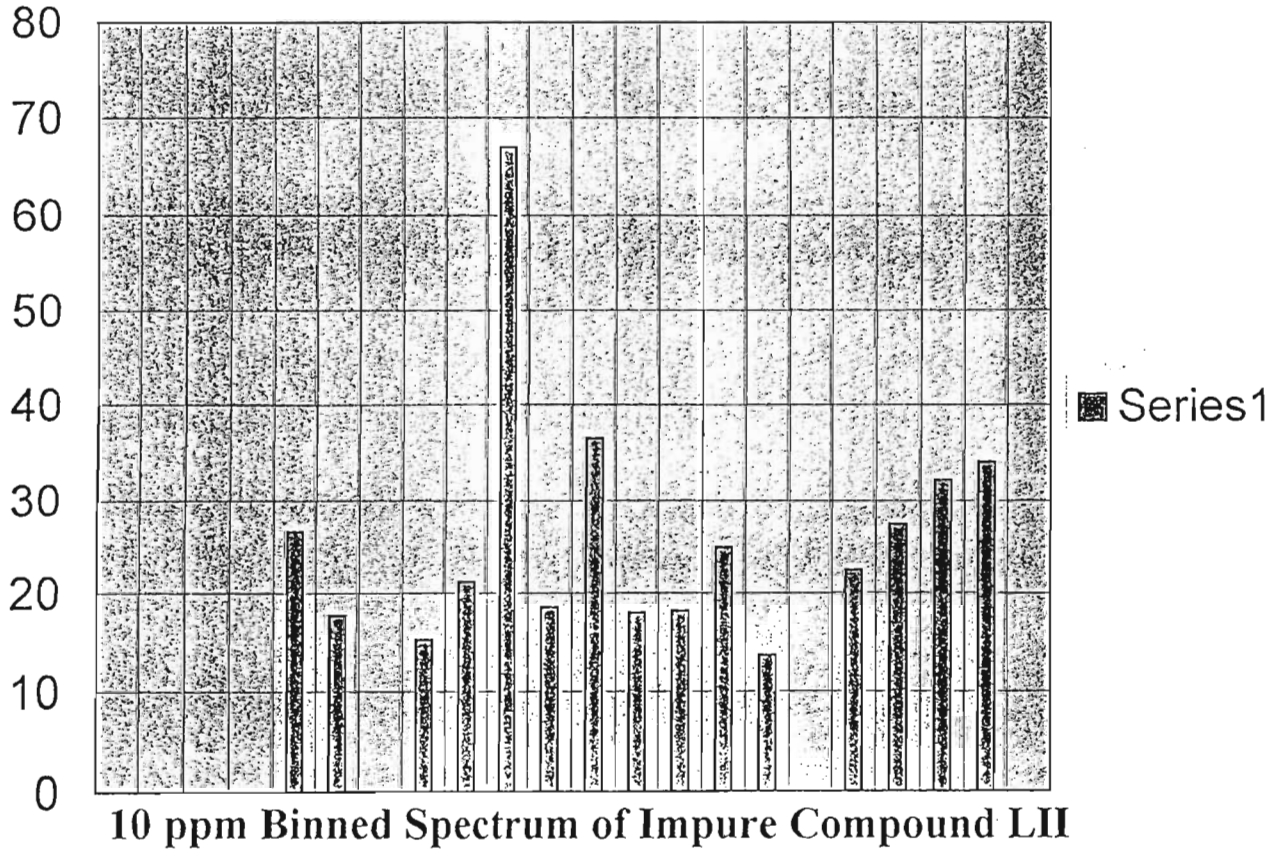
5 ppm Binned Spectrum of Pure Compound LII



5 ppm Binned Spectrum of Impure Compound LII



10 ppm Binned Spectrum of Pure Compound LII



10 ppm Binned Spectrum of Impure Compound LII

^{13}C NMR Data for Impure Compound IX

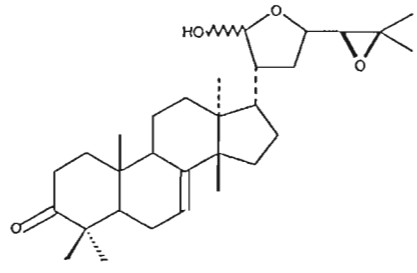
ppm	Intensity
217.315	60.379
•217.257	52.515
145.905	102.02
•145.742	84.835
•145.614	26.2
118.044	73.341
•117.942	62.971
101.63	52.838
97.495	52.982
78.829	19.559
• 78.228	75.864
• 77.935	75.864
• 77.292	75.864
76.655	132.169
• 69.075	20.316
• 68.656	24.444
67.694	66.032

ppm	Intensity
65.326	64.021
• 59.988	20.564
57.708	90.063
57.101	82.725
• 53.472	29.902
• 53.061	19.733
• 52.224	76.332
• 52.17	68.359
50.971	33.921
• 50.823	95.549
50.587	117.17
• 50.445	26.098
• 50.296	50.152
• 49.274	46.854
• 48.698	22.903
• 48.162	73.383
• 48.106	64.979

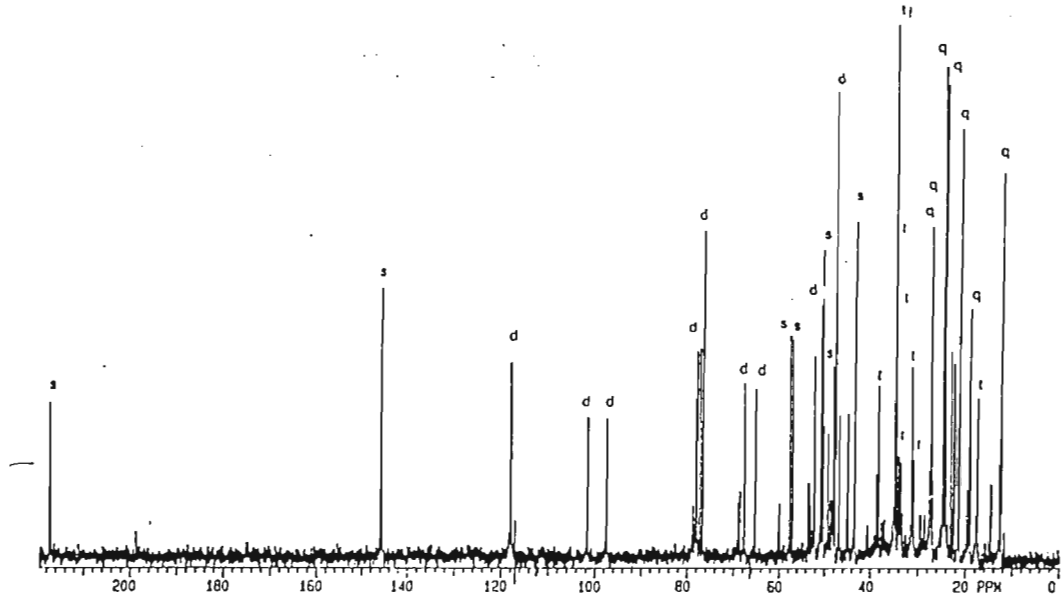
ppm	Intensity
47.657	179.21
46.761	54.423
• 44.951	54.985
• 43.506	99.531
43.33	134.753
• 38.708	31.917
38.271	65.747
• 34.953	49.217
• 34.844	207.354
• 34.654	99.014
34.059	38.353
• 33.754	20.894
33.619	36.123
33.489	20.145
• 33.299	16.783
31.261	72.553
30.915	36.716

ppm	Intensity
27.438	43.325
27.254	128.629
27.039	123.71
24.704	197.419
24.277	184.816
• 24.128	68.46
23.01	85.249
• 22.342	74.247
• 21.471	34.495
21.309	166.123
• 19.777	25.769
• 19.641	23.794
• 19.18	91.208
• 19.012	94.754
17.482	60.523
• 14.592	27.261
• 12.833	36.513
12.461	148.876

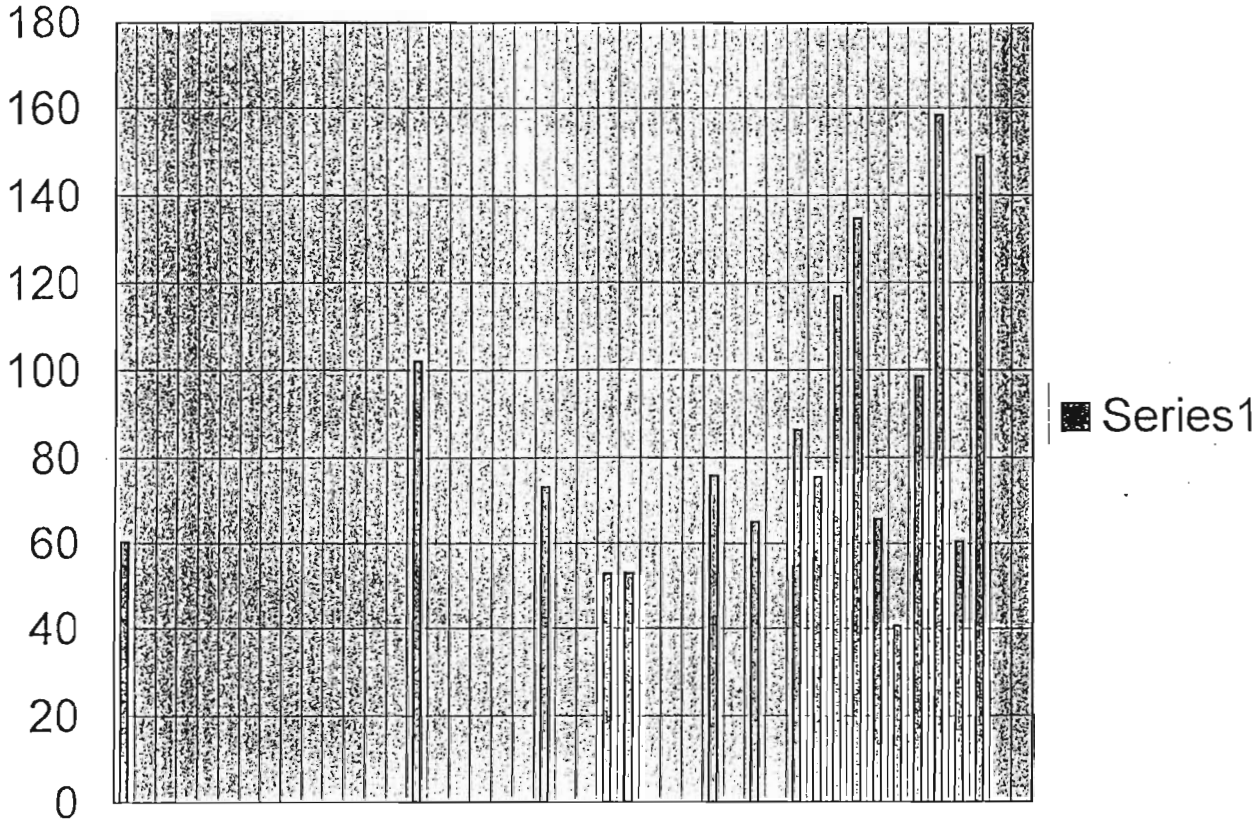
• - denotes extraneous peaks



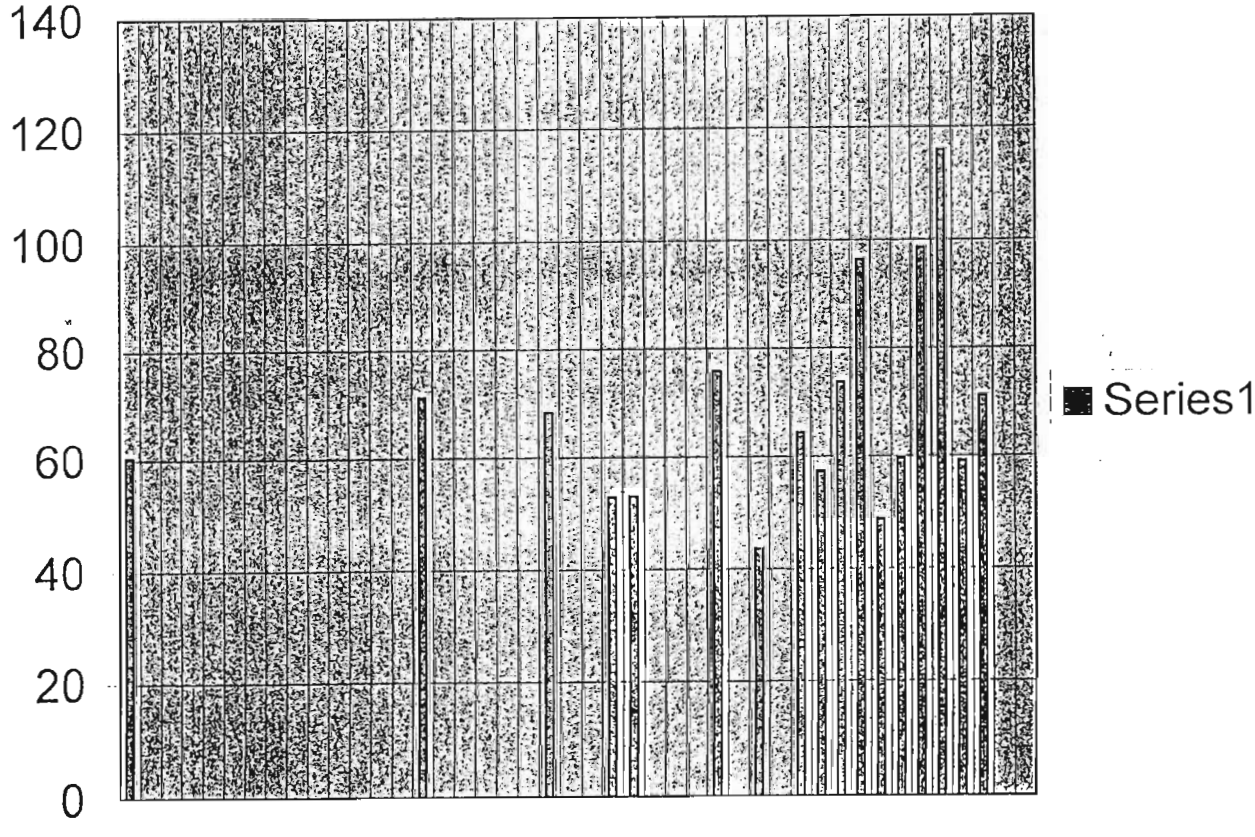
MELIANONE (IX)



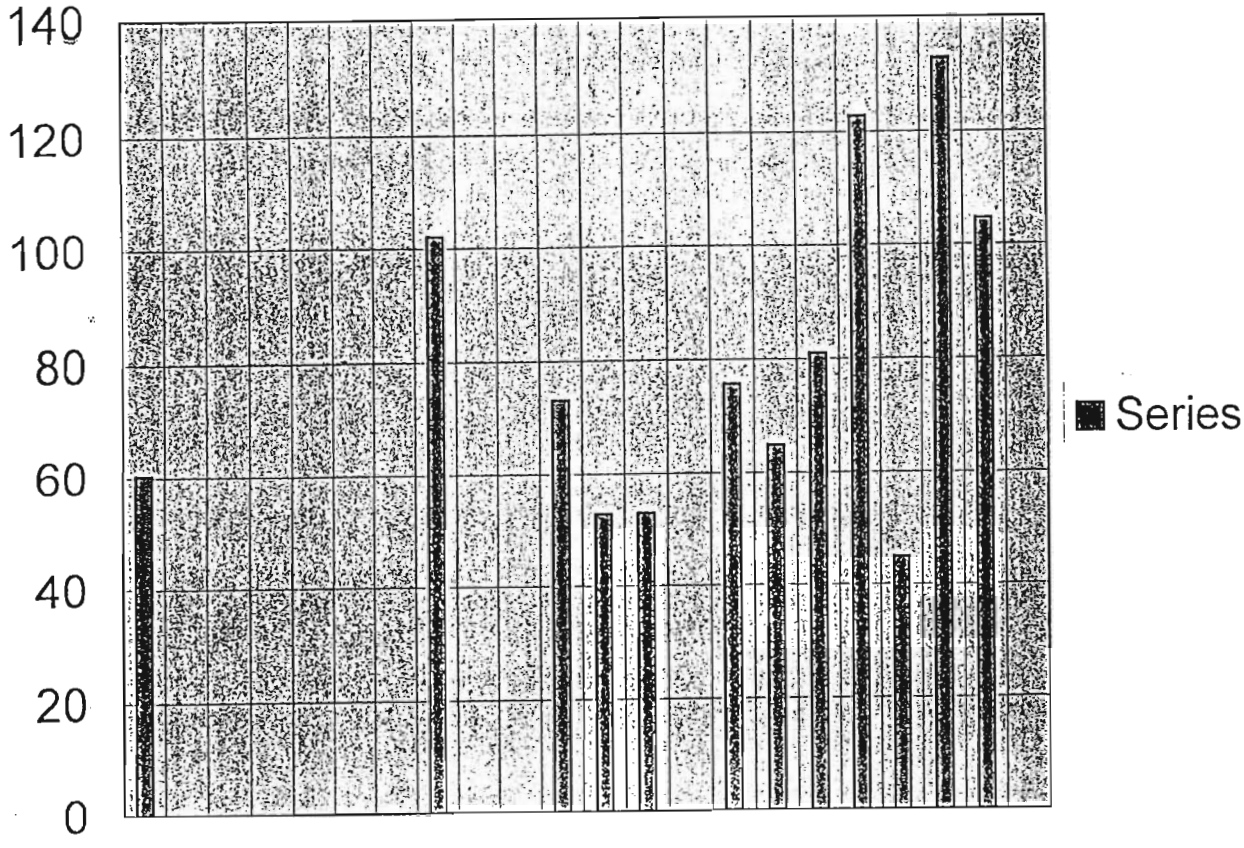
^{13}C NMR Spectrum of Compound IX



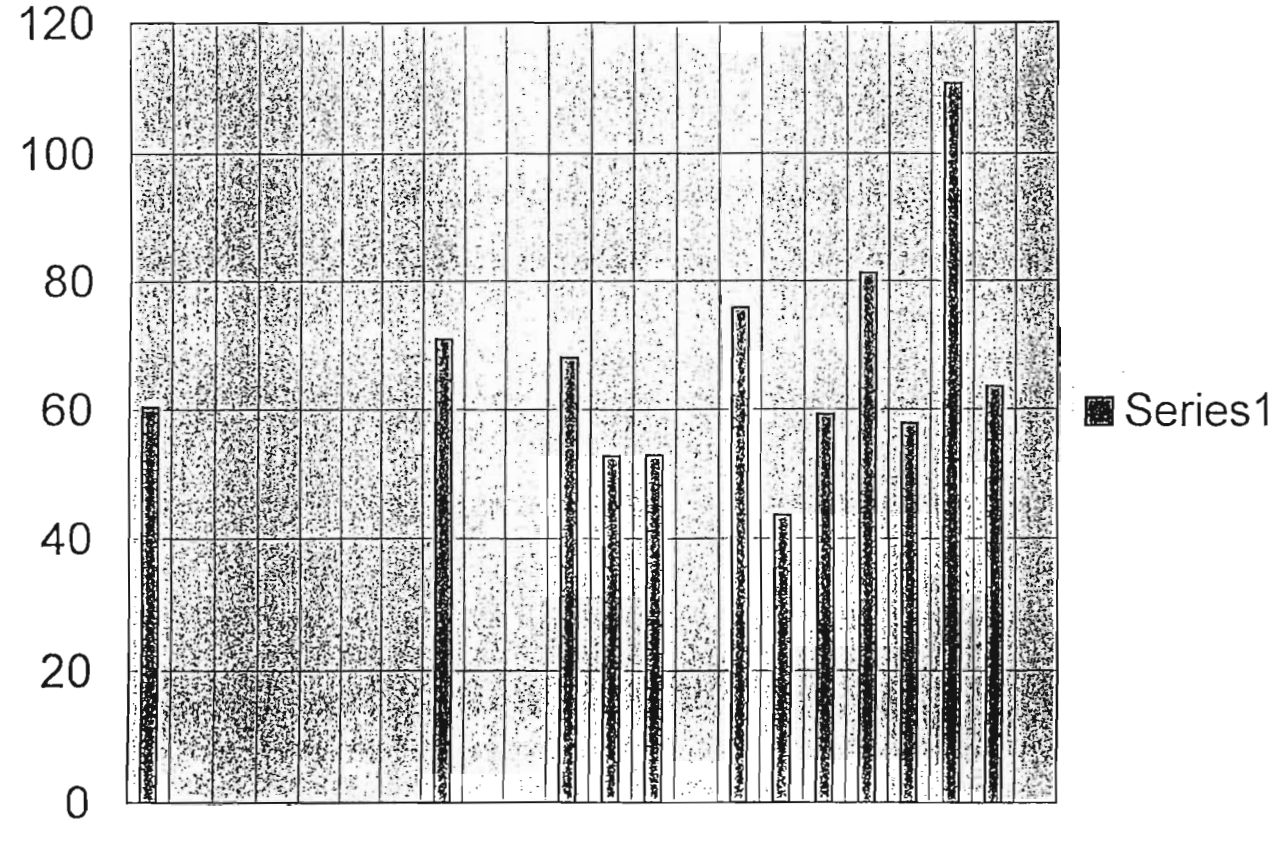
5 ppm Binned Spectrum of Pure Compound IX



5 ppm Binned Spectrum of Impure Compound IX



10 ppm Binned Spectrum of Pure Compound IX



10 ppm Binned Spectrum of Impure Compound IX

^{13}C NMR Data for Impure Compound XIII

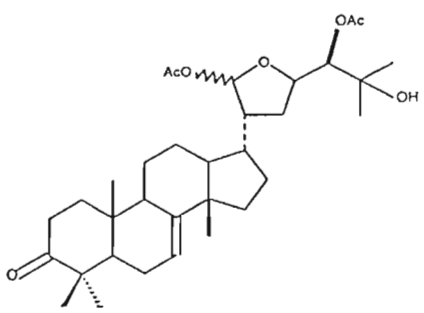
ppm	Intensity
217.61	7.748
•217.515	16.597
•171.806	7.711
171.602	16.086
170.559	18.742
• 146.138	8.366
145.74	23.781
118.751	28.684
• 118.65	6.776
•118.472	9.153
•104.317	9.193
100.372	29.849
• 77.939	19.59863
77.775	15.064
77.468	43.325
• 77.302	19.59863
• 77.063	12.171

ppm	Intensity
• 76.857	7.891
• 76.665	19.59863
76.481	35.854
• 76.354	6.601
• 72.789	9.69
72.648	26.193
52.56	33.208
51.169	27.42
• 50.983	12.59
50.121	29.906
• 48.548	16.091
48.447	32.669
48.057	27.962
46.751	33.117
• 46.488	9.672
• 44.782	9.14
• 43.89	6.172

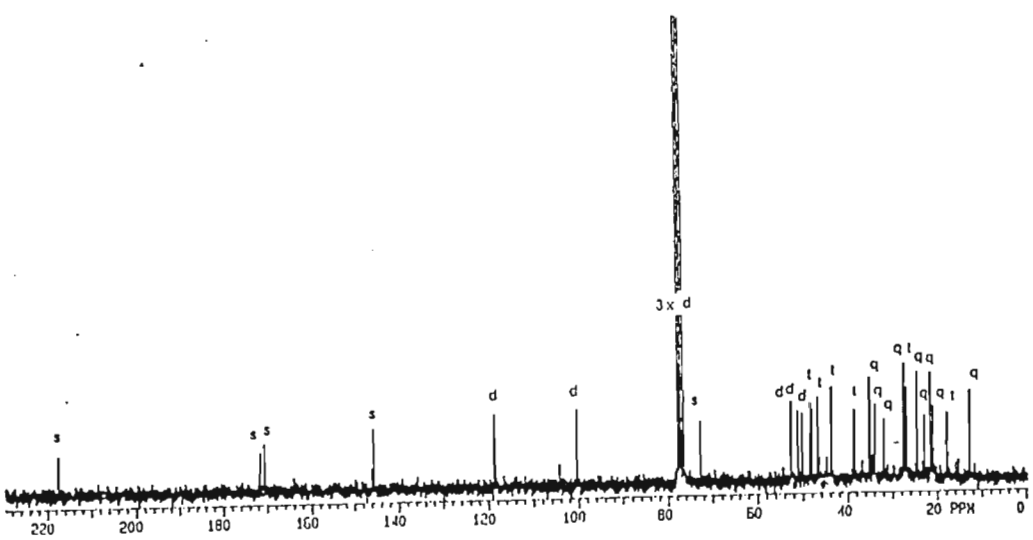
ppm	Intensity
43.744	36.246
38.617	27.855
• 36.708	7.6
35.23	41.043
35.04	31.574
• 34.399	10.616
33.932	29.613
33.873	29.803
• 32.084	9.52
31.882	25.679
27.62	43.544
27.458	44.435
27.409	39.303
• 27.308	8.178
• 26.94	6.256
26.857	35.585
• 24.951	6.816

ppm	Intensity
24.627	44.408
24.476	35.221
• 23.4	6.75
22.919	27.403
21.675	42.553
• 21.556	7.818
21.461	23.498
21.04	30.213
• 17.854	25.826
• 15.955	6.146
• 15.33	7.95
12.817	36.871

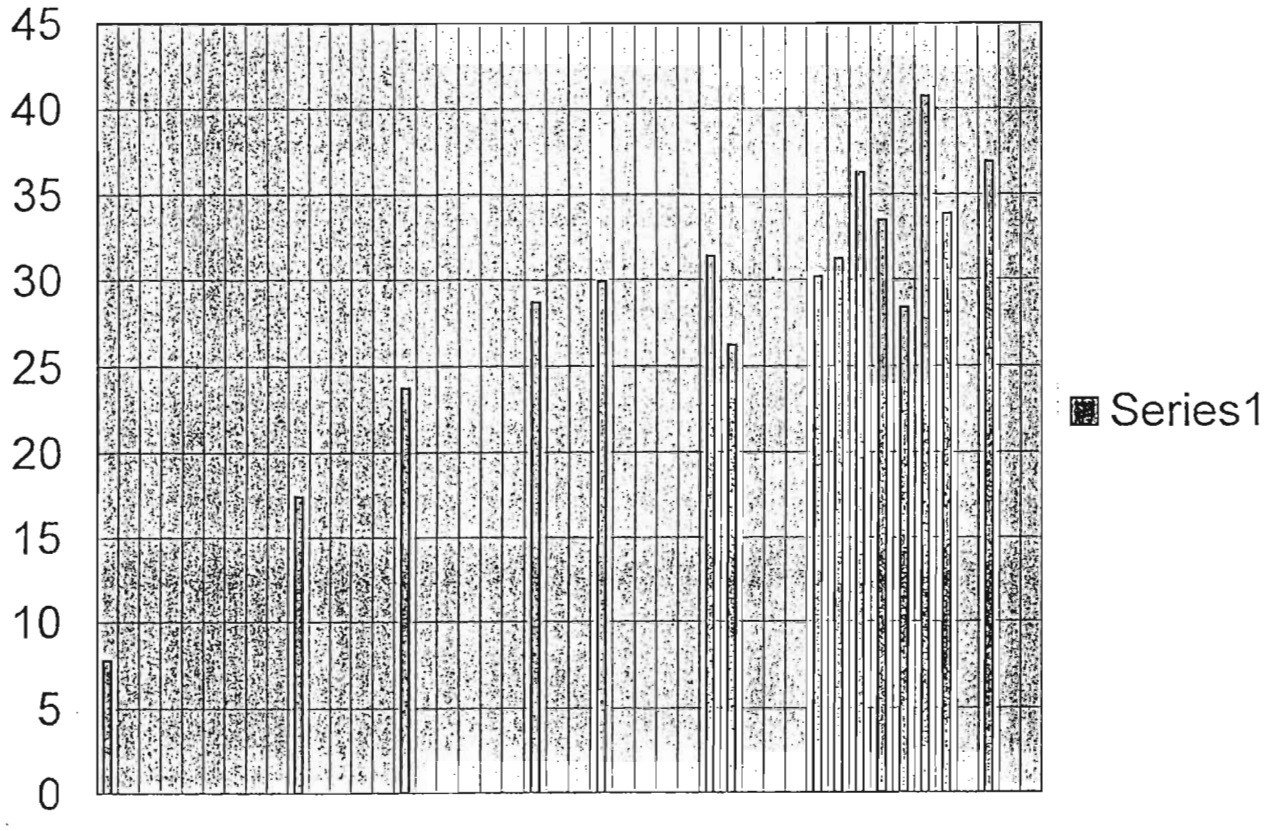
• - denotes extraneous peaks



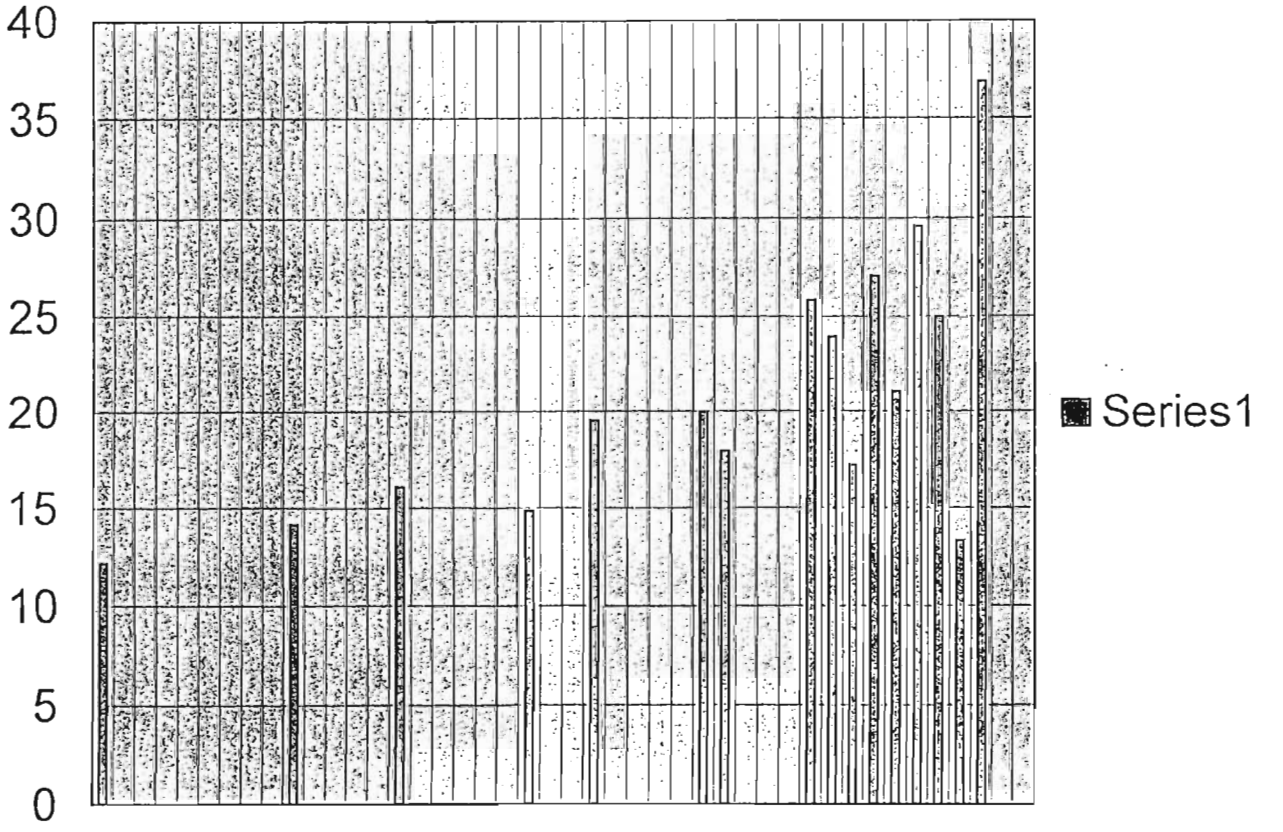
MELIANODIOL ACETATE (XIII)



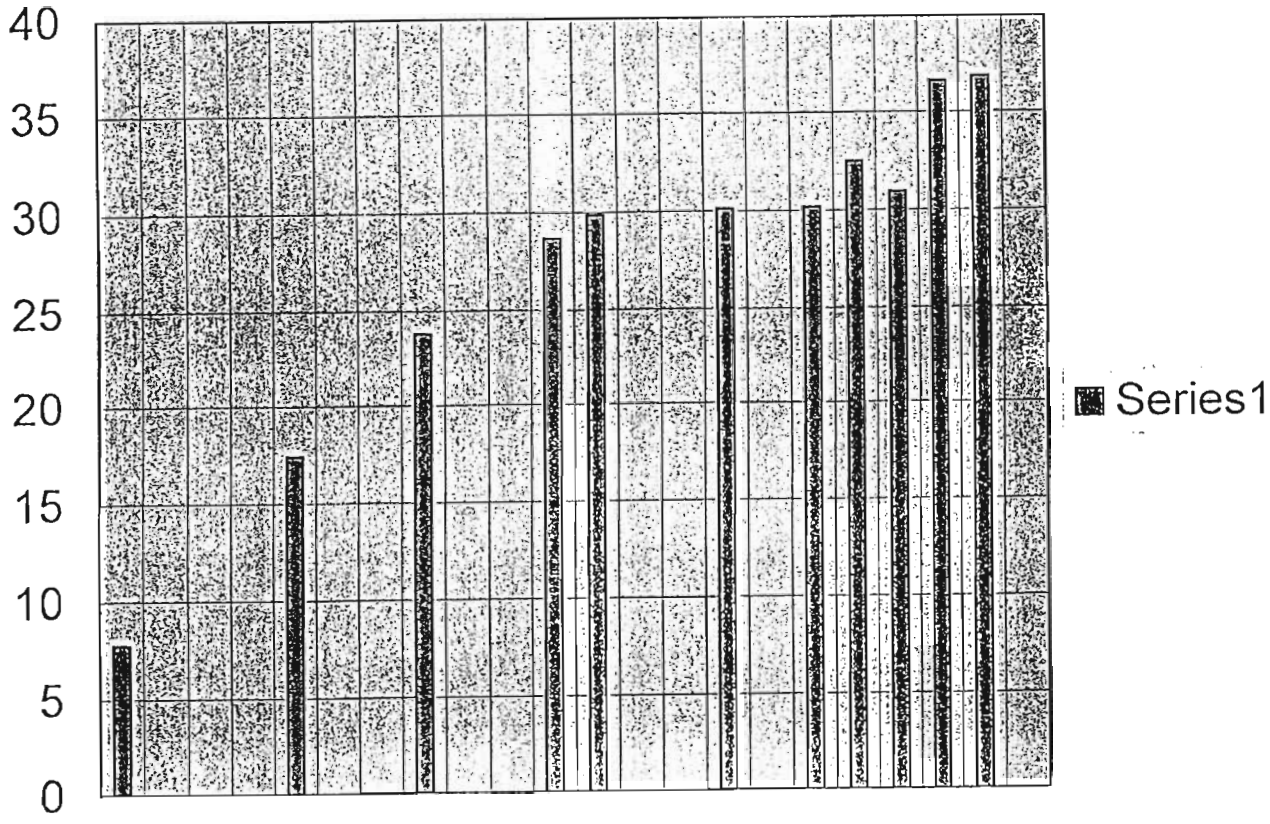
^{13}C NMR Spectrum of Compound XIII



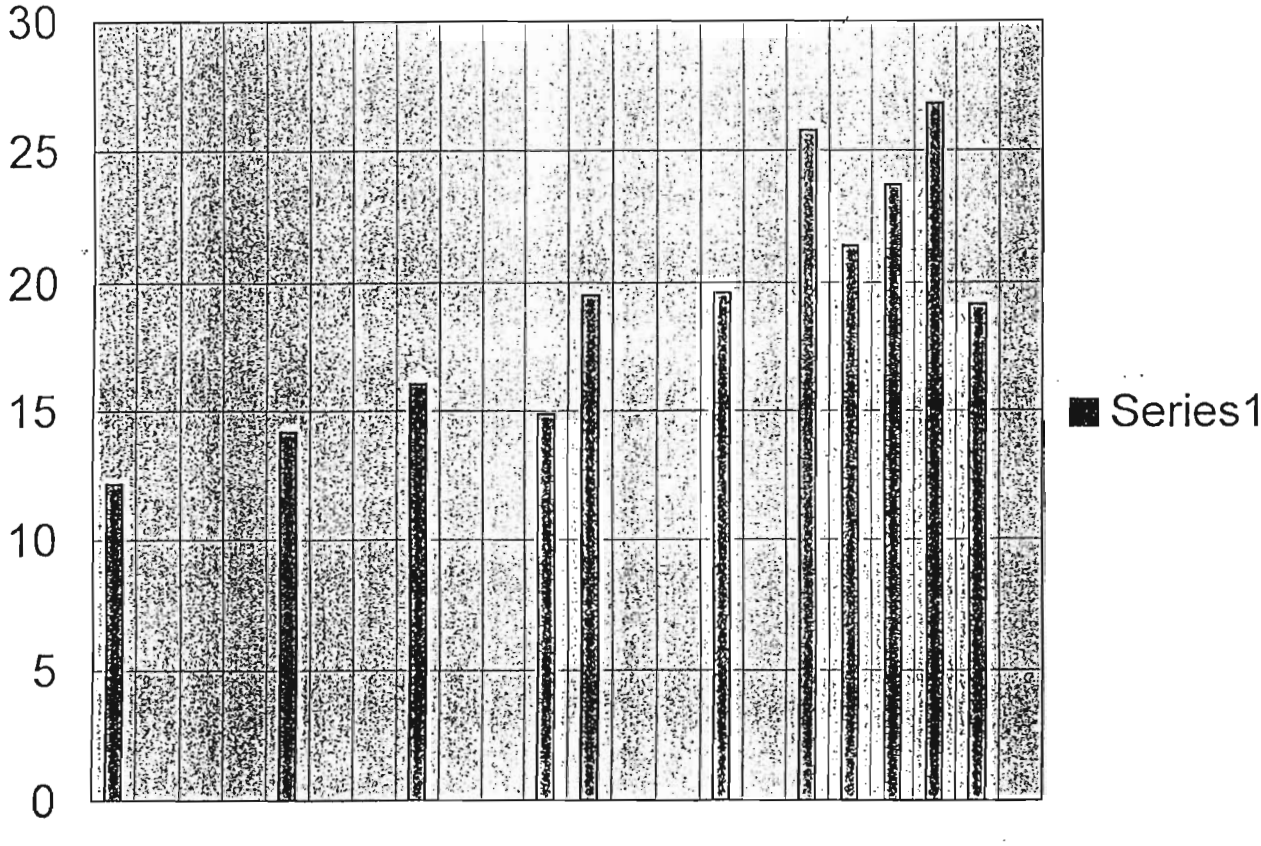
5 ppm Binned Spectrum of Pure Compound XIII



5 ppm Binned Spectrum of Impure Compound XIII



10 ppm Binned Spectrum of Pure Compound XIII



10 ppm Binned Spectrum of Impure Compound XIII

^{13}C NMR Data for Impure Compound XXXXV

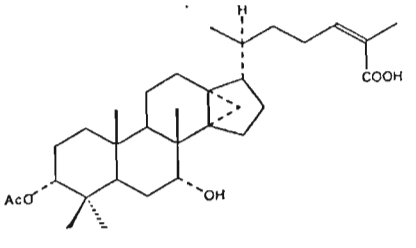
ppm	Intensity
173.002	64.227
• 171.121	26.248
171.076	61.382
• 162.286	26.282
146.838	92.814
• 146.756	34.283
126.059	101.577
• 119.69	27.87
78.138	115.701
• 77.245	11.104
74.404	99.965
• 72.327	29.177
• 60.228	33.884
52.197	108.292
• 46.986	31.689
• 44.325	30.829
44.041	121.236

ppm	Intensity
• 41.868	36.262
• 41.643	32.837
41.177	110.759
38.871	111.998
• 38.123	10.972
• 37.508	36.627
37.193	121.362
• 37.085	15.785
36.168	132.089
36.139	72.027
35.438	122.641
• 35.256	29.518
• 35.033	20.118
34.942	125.139
• 33.944	15.024
• 33.857	32.796
33.755	88.92

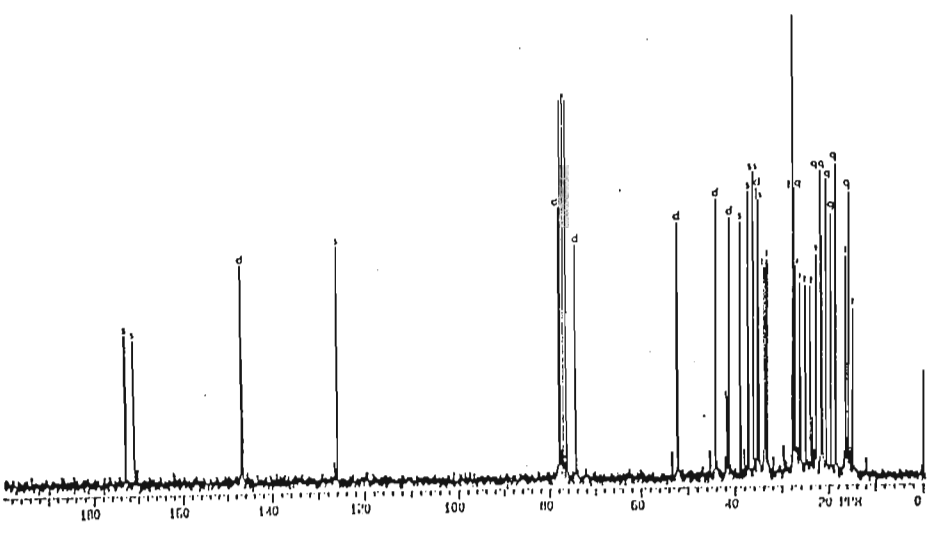
ppm	Intensity
• 33.589	38.781
33.189	92.262
• 29.695	12.025
• 27.9	43.553
27.588	198.182
27.453	125.867
• 27.22	12.7
27.09	89.819
• 26.783	11.076
• 26.54	27.59
• 26.422	12.096
26.103	81.871
25.066	80.632
24.056	80.363
• 23.519	23.937
22.783	93.586
21.827	129.557

ppm	Intensity
• 21.763	50.914
21.473	102.036
• 21.299	16.458
20.587	126.927
• 19.677	17.46
19.538	117.913
• 18.627	45.949
18.522	132.843
• 16.496	28.995
16.354	94.107
• 15.86	15.614
15.711	121.722
• 15.158	35.717
14.801	70.733
• 0.001	43.636

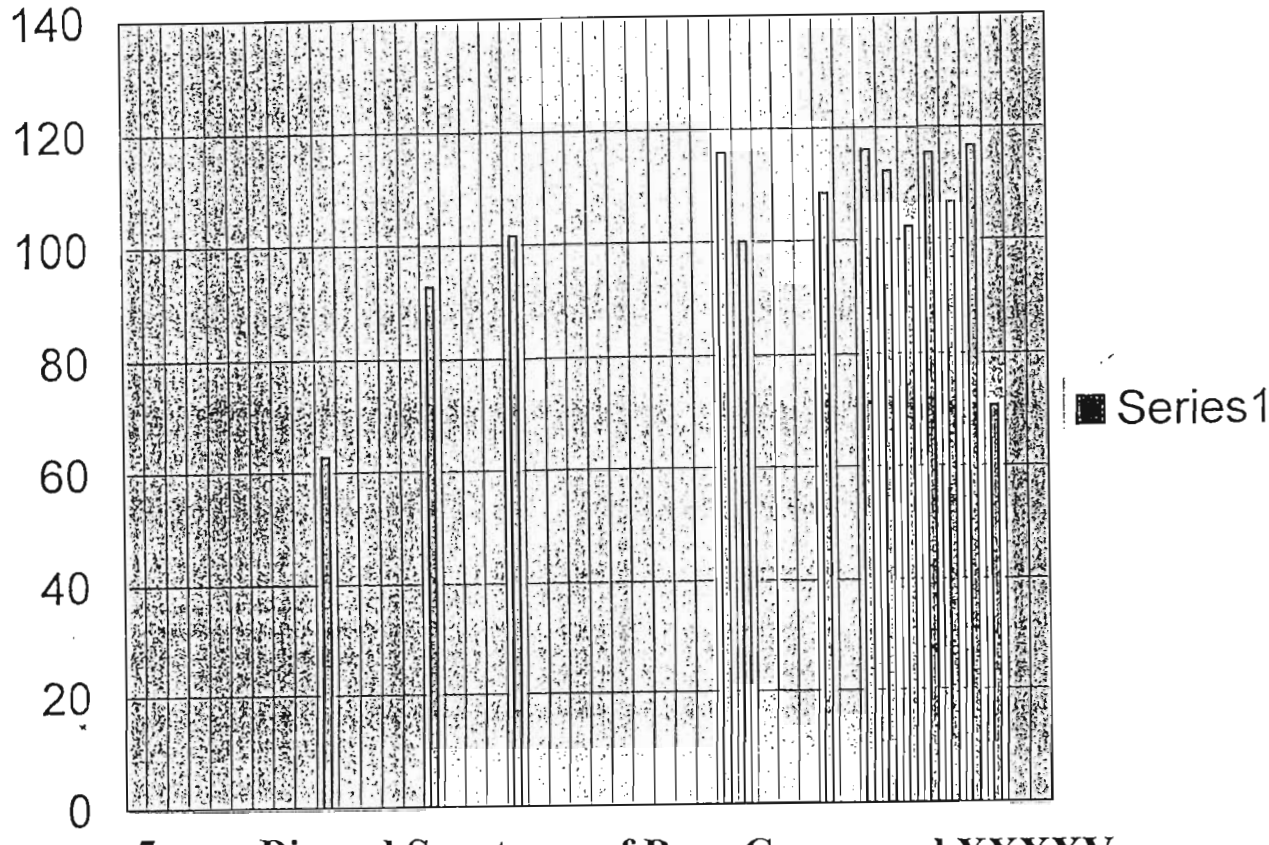
• - denotes extraneous peaks



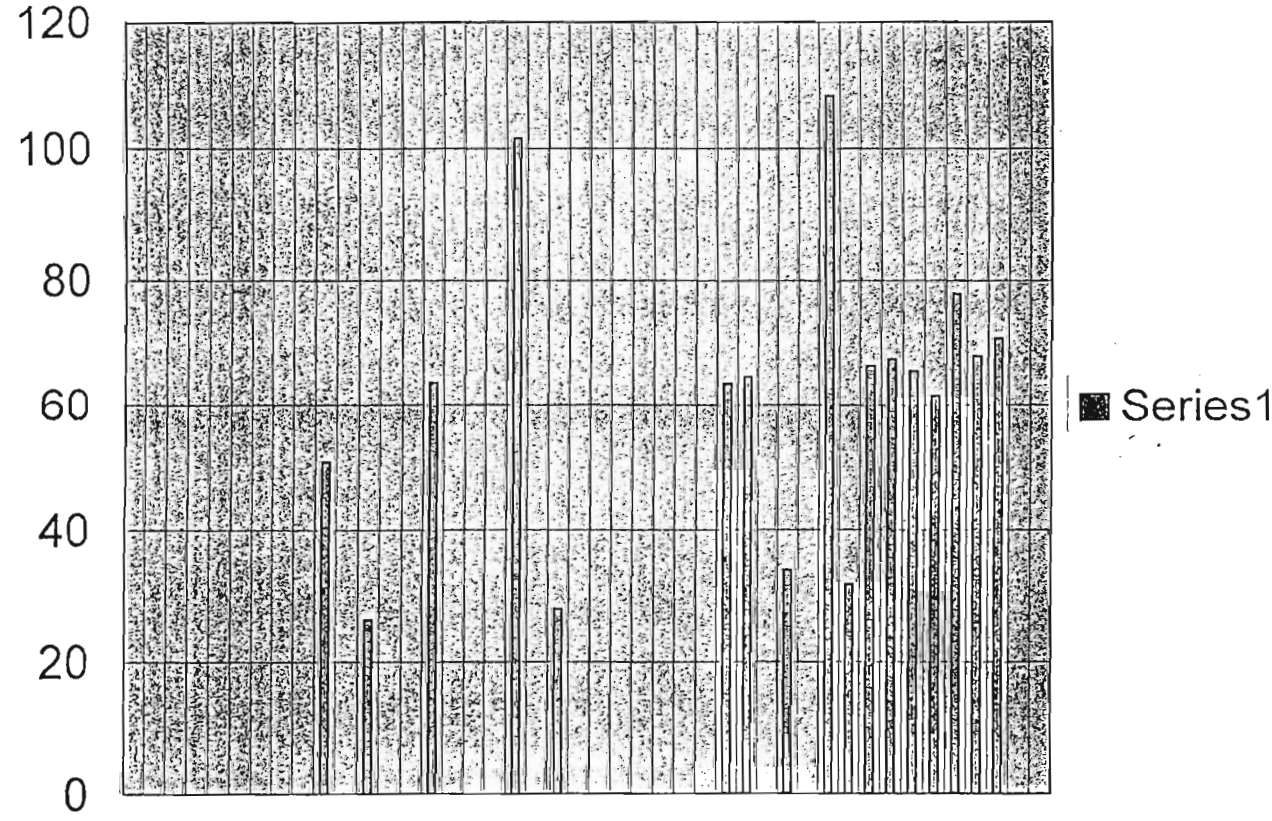
COMPOUND XXXXV



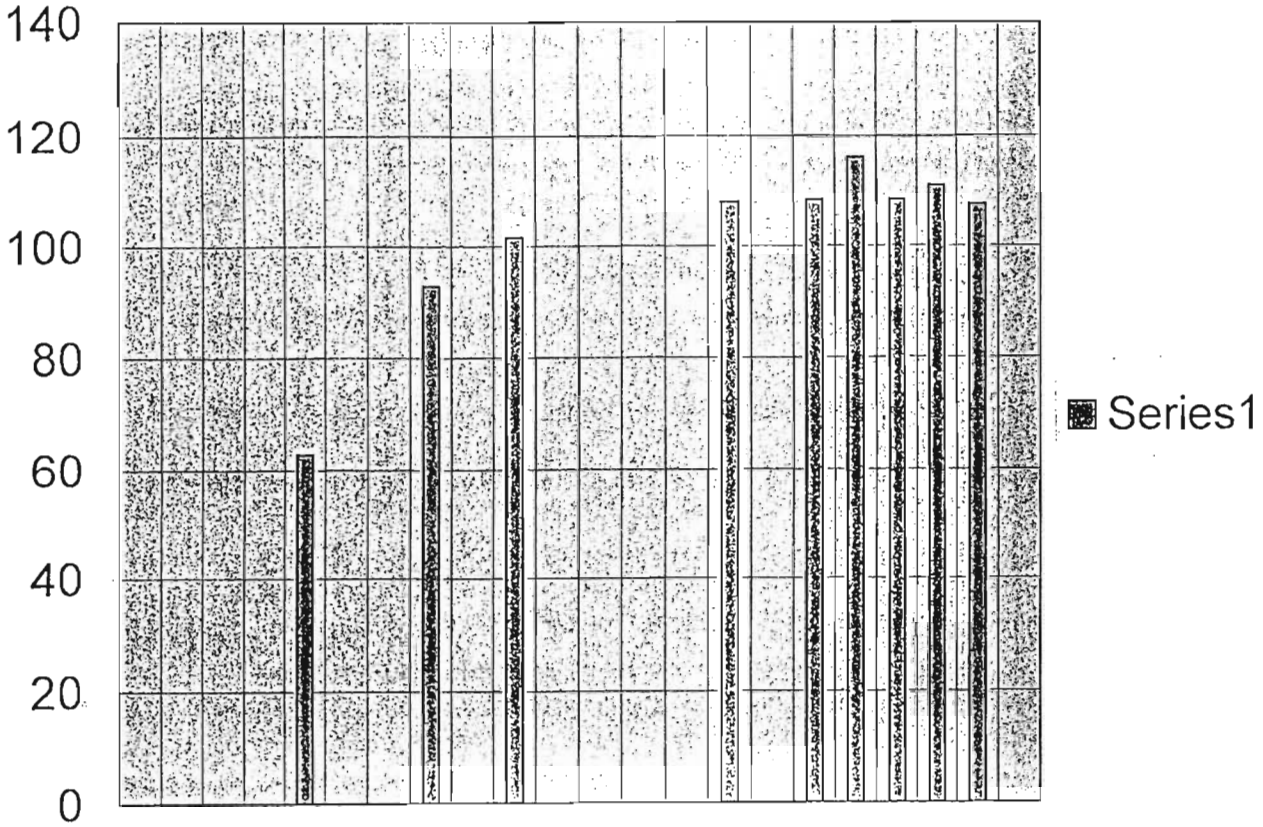
^{13}C NMR Spectrum of Compound XXXXV



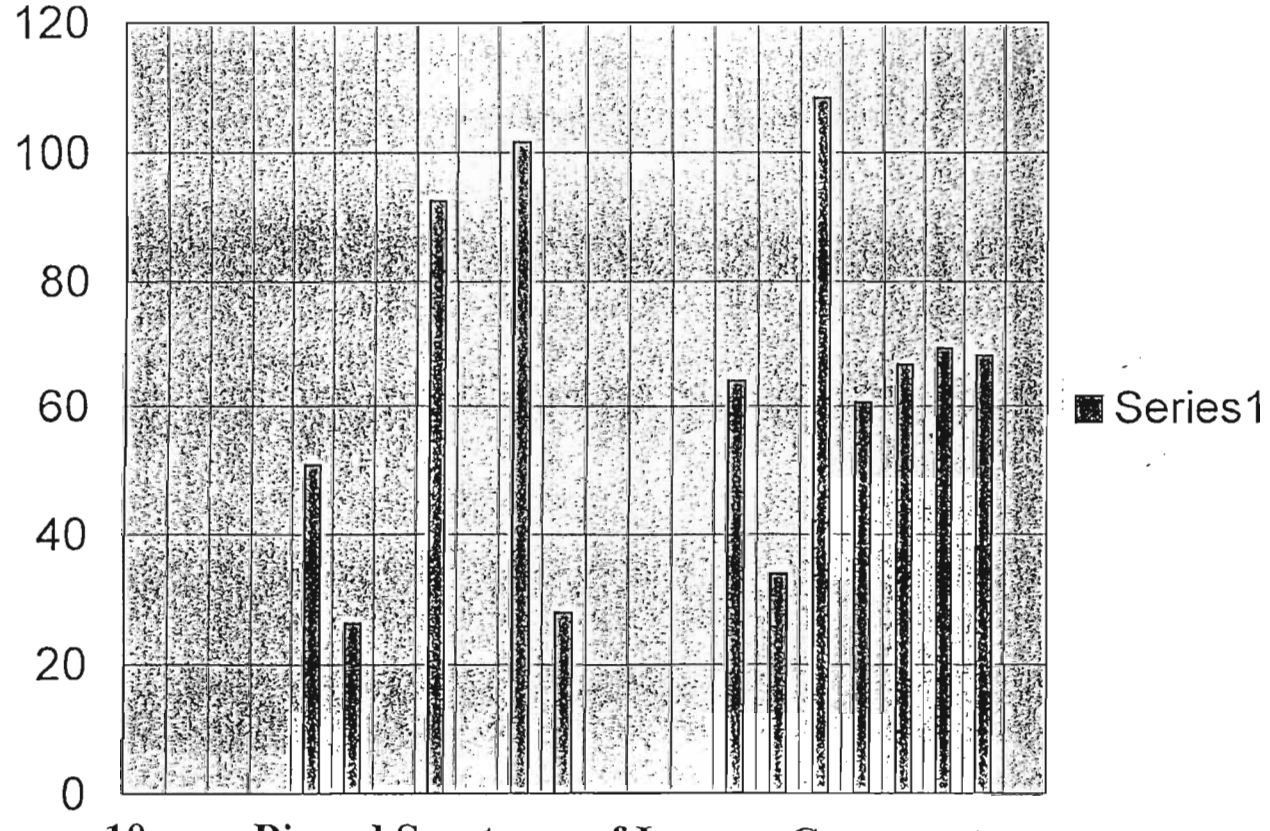
5 ppm Binned Spectrum of Pure Compound XXXXV



5 ppm Binned Spectrum of Impure Compound XXXXV



10 ppm Binned Spectrum of Pure Compound XXXXV



10 ppm Binned Spectrum of Impure Compound XXXXV

^{13}C NMR Data for Impure Compound LIV

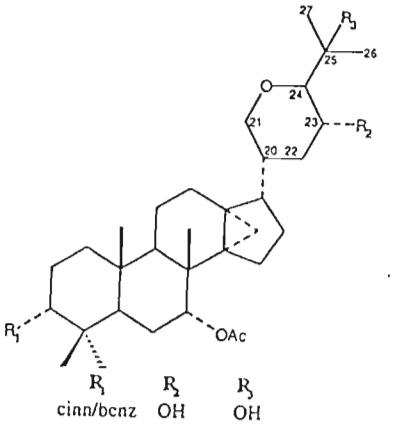
ppm	Intensity
169.826	10.973
166.5	8.986
165.612	16.833
144.3	7.632
134.5	8.5
132.818	31.147
131.087	17.127
130.265	12.389
129.417	65.085
128.939	32.914
•128.382	60.043
•127.979	29.679
118.93	18.181
86.647	32.042
78.946	31.596
• 78.255	14.521
• 78.06	10.613
• 78.014	10.591

ppm	Intensity
• 77.881	11.635
• 77.22	137.594
76.325	34.952
• 76.255	23.961
74.22	32.141
70.775	22.108
64.931	30.704
45.947	29.779
45.525	20.667
45.388	14.54
• 42.498	19.286
42.347	20.934
• 40.548	23.754
• 40.502	17.615
38.249	27.991
37.343	22.95
• 37.25	14.157
37.141	31.93

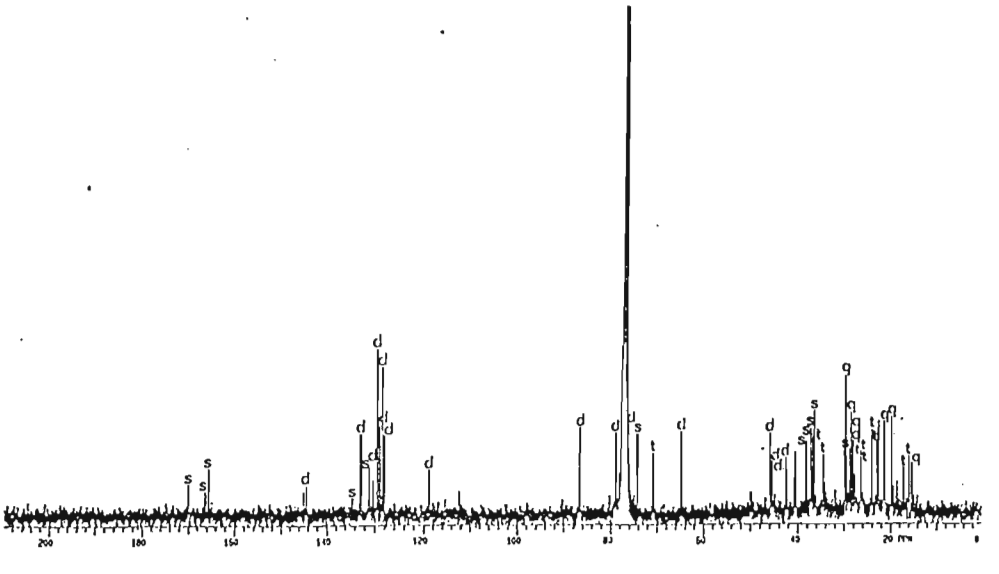
ppm	Intensity
36.809	30.456
36.461	41.508
34.621	21.186
• 34.245	11.167
• 29.805	12.107
29.717	53.475
28.993	17.253
28.883	23.462
• 28.712	21.598
• 28.579	38.136
• 28.273	26.976
28.02	28.24
27.75	13.66
27.708	10.426
26.446	23.734
26.289	12.24
24.138	31.177
23.084	25.825

ppm	Intensity
• 22.821	12.608
• 22.717	34.347
• 21.581	27.52
21.5	34.816
• 21.367	27.353
20.001	36.104
• 18.746	11.48
17.225	16.383
16.089	22.416
15.989	17.096
• 15.404	18.748

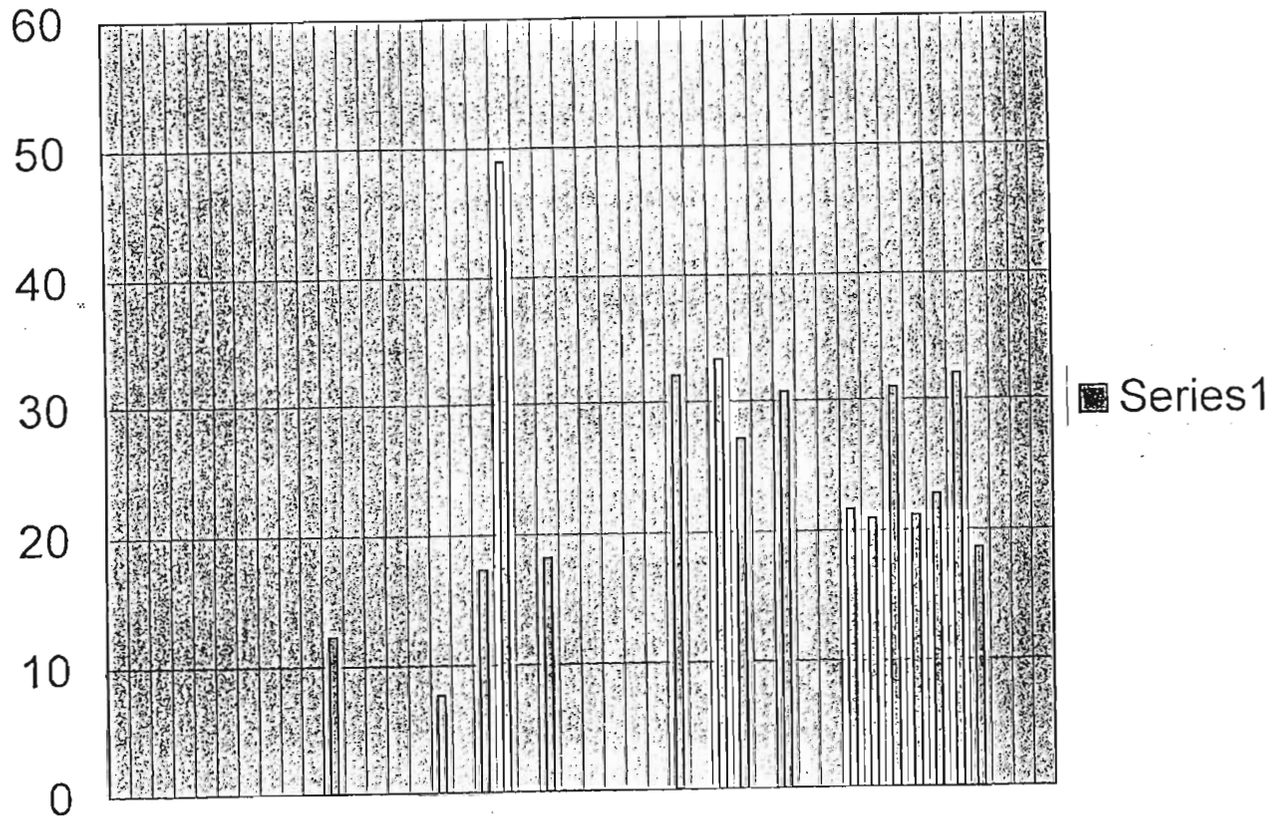
• - denotes extraneous peaks



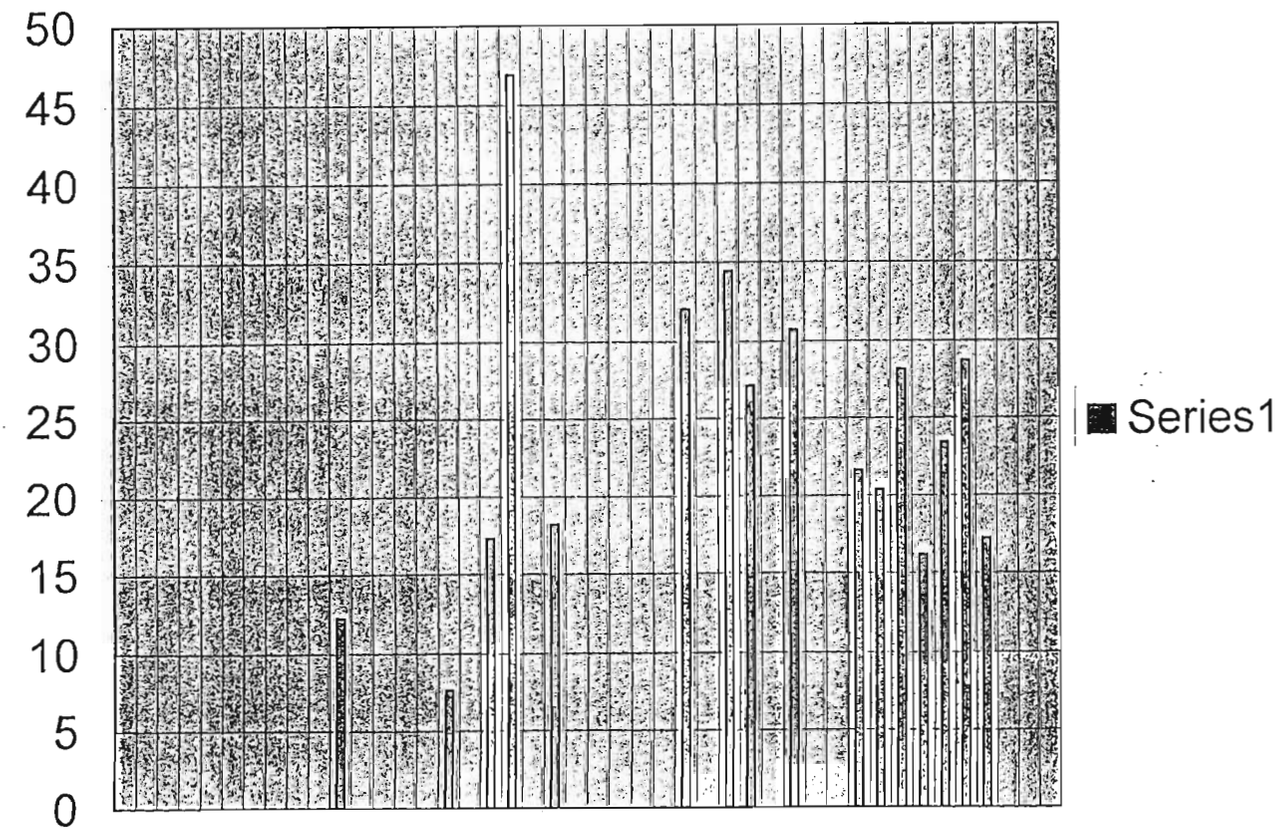
COMPOUND LIV



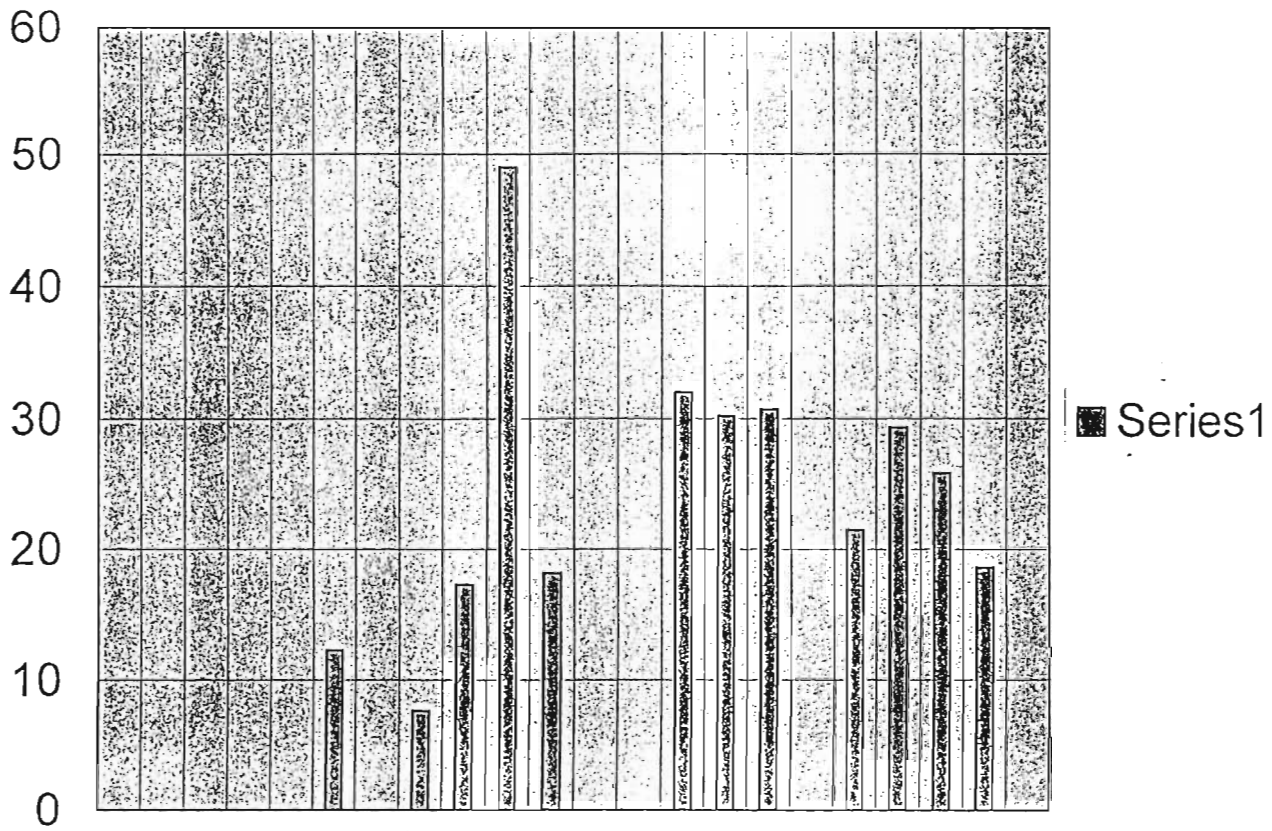
^{13}C NMR Spectrum of Compound LIV



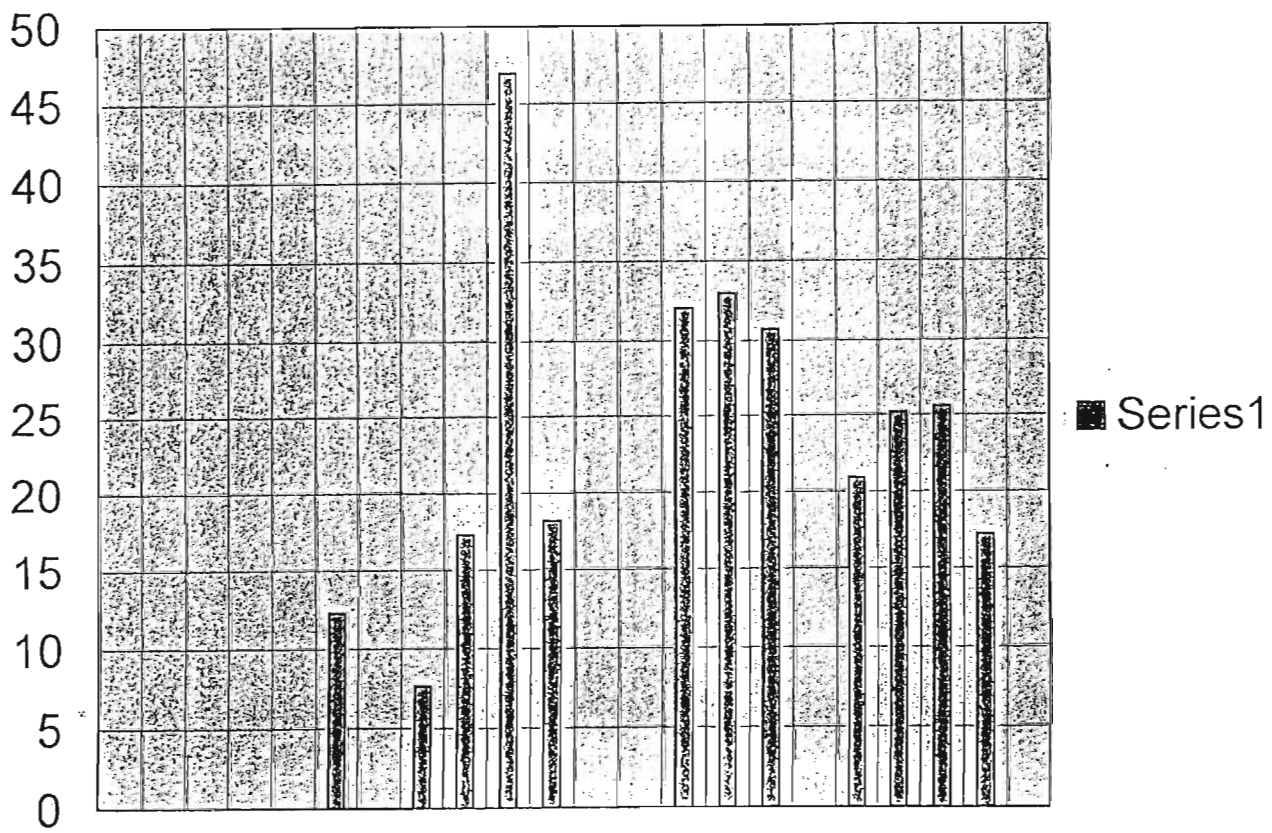
5 ppm Binned Spectrum of Pure Compound LIV



5 ppm Binned Spectrum of Impure Compound LIV



10 ppm Binned Spectrum of Pure Compound LIV



10 ppm Binned Spectrum of Impure Compound LIV

^{13}C NMR Data for Impure Compound LI

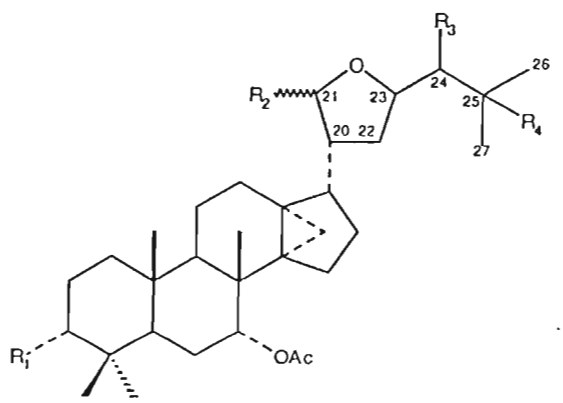
ppm	Intensity
169.843	12.746
165.603	10.89
144.3	5.75
134.5	5.86
• 132.821	10.196
132.786	22.438
129.394	52.834
128.905	17.328
• 128.35	50.345
• 127.957	12.47
118.8	5.76
97.781	14.905
• 97.76	14.382
78.95	21.495
78.668	29.96
• 77.197	36.836
• 76.347	20.746

ppm	Intensity
75.164	20.388
• 73.329	11.671
73.296	11.538
48.961	23.125
45.564	20.922
44.897	24.551
42.431	23.84
• 38.23	11.902
38.193	22.259
37.22	28.093
37.167	25.518
36.781	28.703
34.531	18.168
• 34.154	10.612
• 31.91	14.732
• 29.683	71.763
29.585	13.614

ppm	Intensity
29.437	27.992
• 29.343	16.785
• 29.239	10.91
29.139	11.658
• 29.083	13.313
28.979	21.126
28.002	27.805
27.464	20.869
26.701	39.201
26.636	37.576
26.518	26.304
• 26.478	15.803
• 26.457	14.522
26.269	24.673
• 23.031	21.69
22.678	29.546
• 21.566	10.726

ppm	Intensity
• 21.541	13.006
21.481	33.965
• 21.331	20.022
19.622	30.149
• 19.578	10.536
16.739	17.004
15.959	27.605
• 15.009	11.897

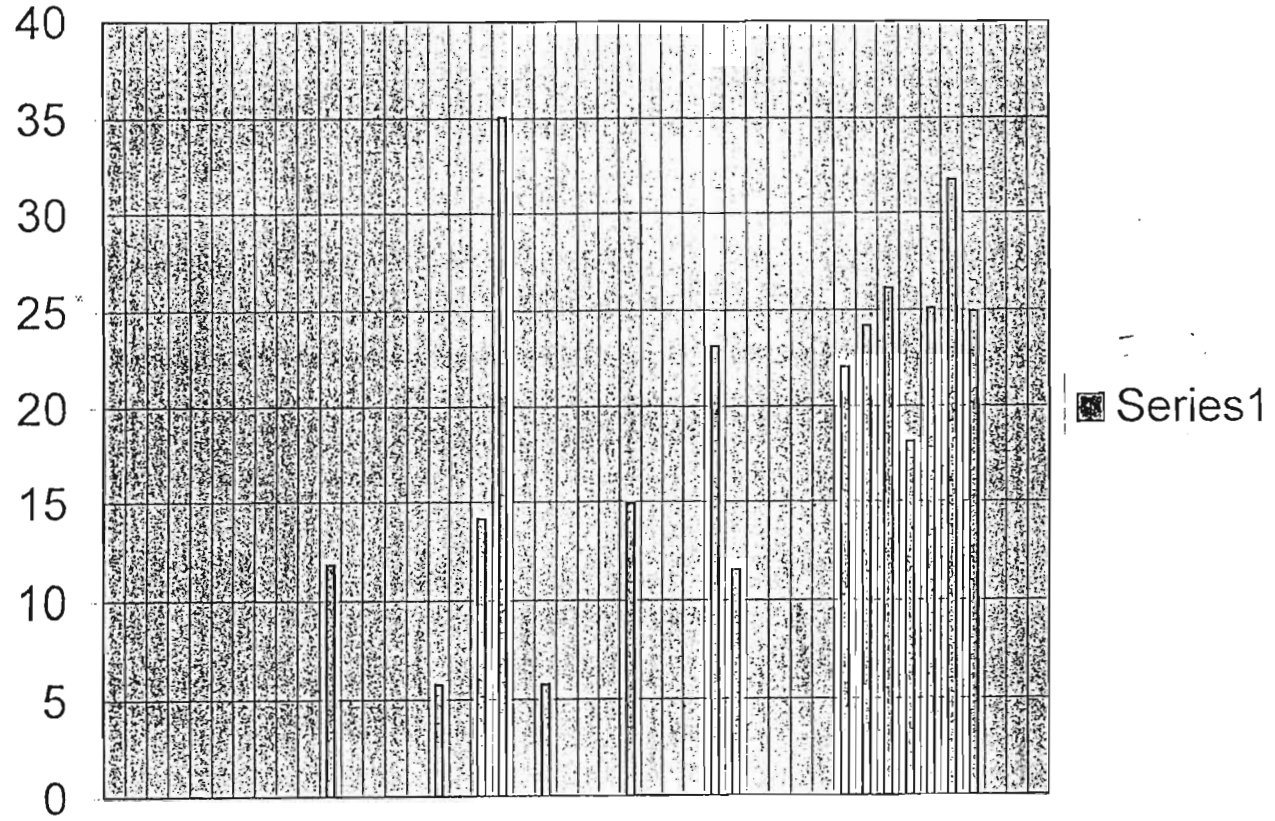
• - denotes extraneous peaks



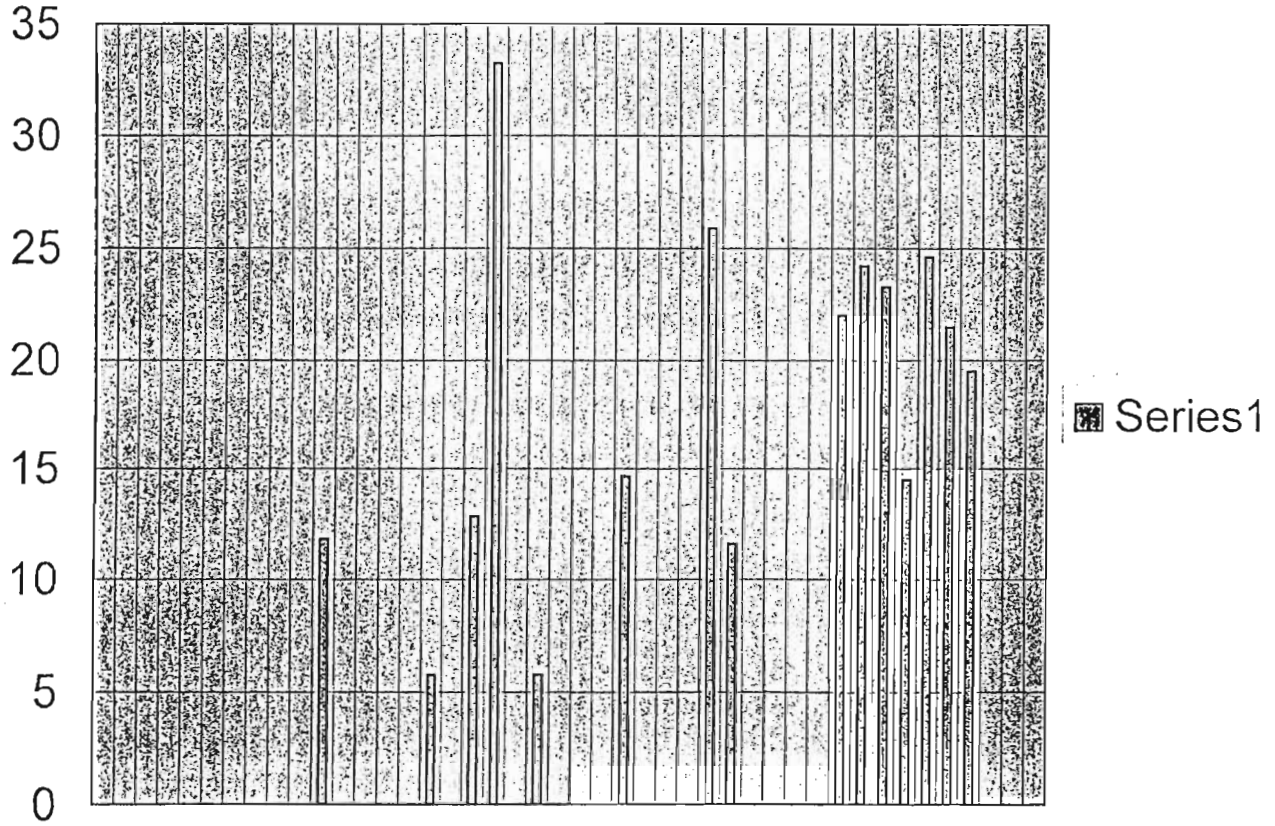
R_1	R_2	R_3	R_4
cinn/benz	OH	OH	OH

COMPOUND LI

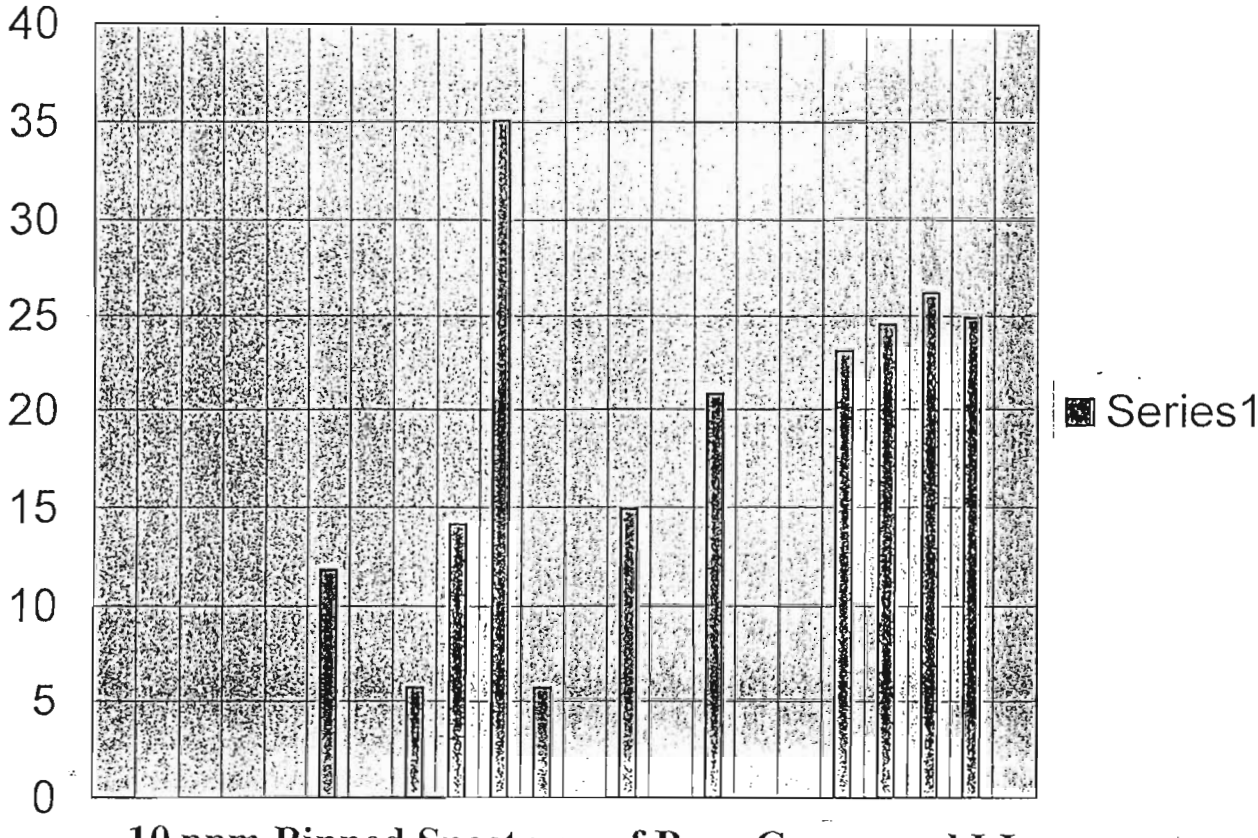
^{13}C NMR Spectrum of Compound LI



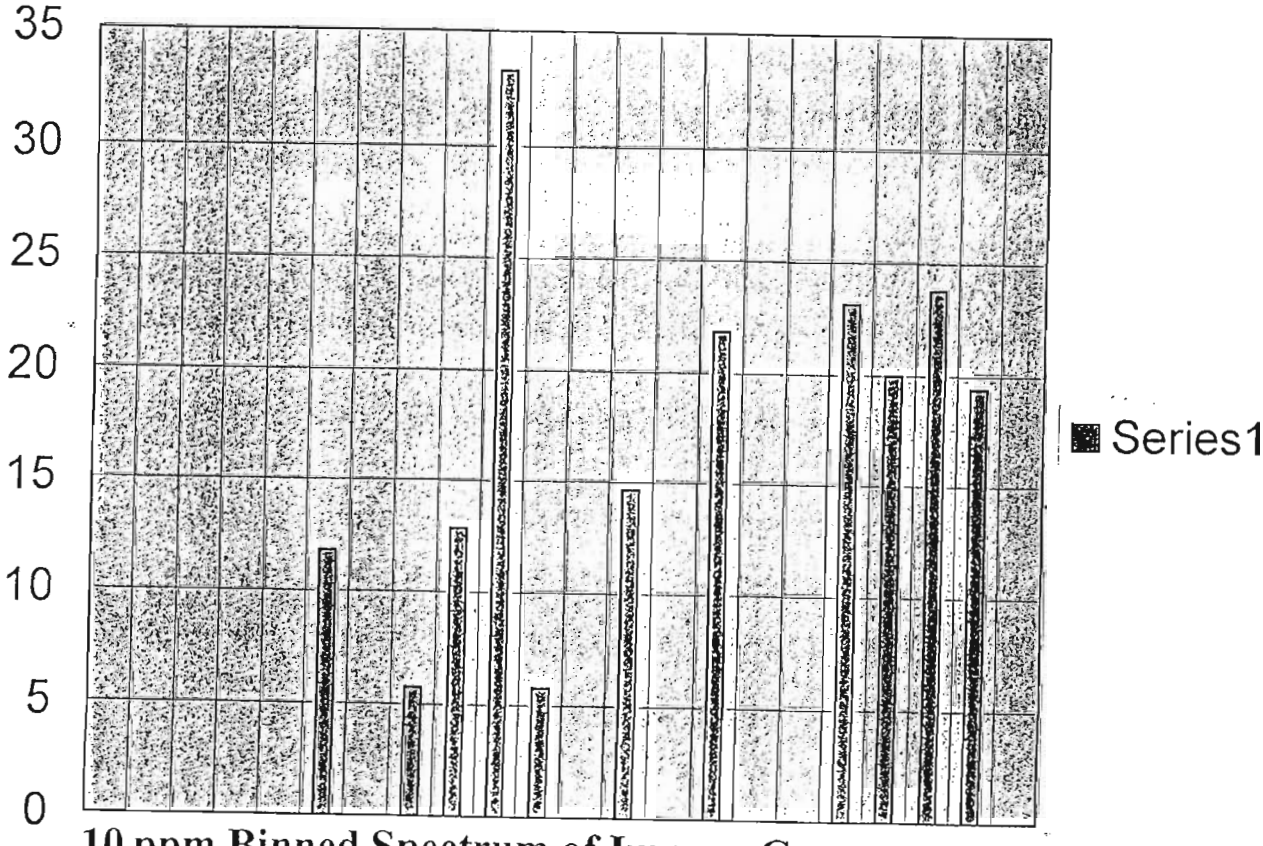
5 ppm Binned Spectrum of Pure Compound LI



5 ppm Binned Spectrum of Impure Compound LI



10 ppm Binned Spectrum of Pure Compound LI

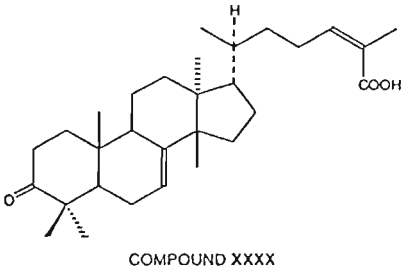


10 ppm Binned Spectrum of Impure Compound LI

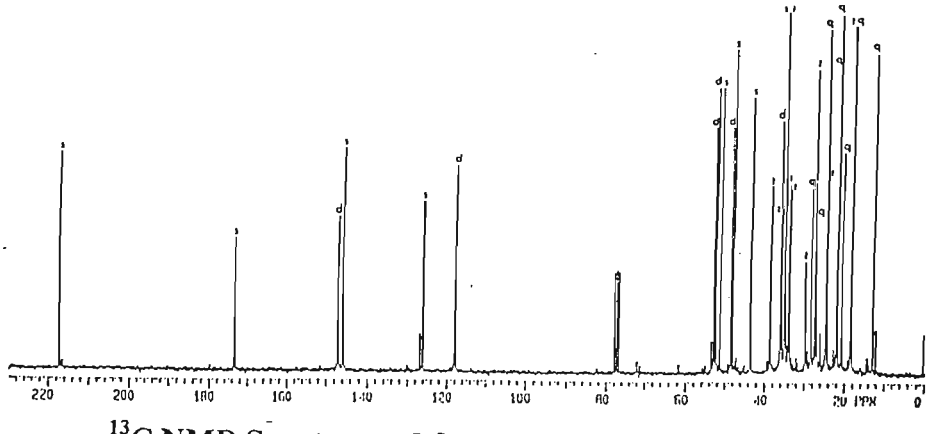
^{13}C NMR Data for Impure Compound XXXX

ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity	ppm	Intensity
217.104	93.987	51.101	131.15	• 34.574	24.609	26.849	68.089	• 14.129	6.346
• 173.332	16.423	48.383	108.351	33.982	82.565	• 26.658	10.383	12.738	140.615
173.199	57.49	47.805	145.021	33.575	80.367	• 25.93	16.692	• 11.928	18.397
146.965	66.514	• 46.899	6.793	• 29.677	48.375	• 24.795	7.388	• -0.003	17.142
145.848	97.991	43.435	121.039	• 29.586	10.639	• 24.668	10.575		
• 145.728	25.352	38.466	81.824	• 29.424	7.671	24.485	150.993		
• 145.511	20.071	36.006	112.369	• 29.337	8.78	24.311	86.18		
• 126.725	15.372	• 35.827	11.114	• 29.28	7.444	• 22.775	6.364		
125.93	73.121	35.586	71.854	• 29.246	6.727	• 22.66	10.037		
• 117.932	6.835	• 35.397	7.396	• 29.042	9.303	• 22.111	9.096		
• 117.809	30.885	• 35.367	6.59	• 28.976	6.656	21.92	136.035		
117.714	89.499	• 35.254	6.469	28.152	81.08	• 21.747	7.114		
• 53.454	12.744	• 35.227	6.44	27.376	133.91	21.545	157.735		
52.807	107.678	• 35.187	6.831	• 27.247	9.83	20.541	98.713		
• 52.62	6.988	• 35.108	7.66	• 27.156	7.769	• 18.559	6.493		
52.252	125.071	34.934	163.663	• 27.1	7.24	• 18.47	8.631		
• 51.242	7.102	• 34.83	87.793	• 27.024	8.143	18.194	154.799		

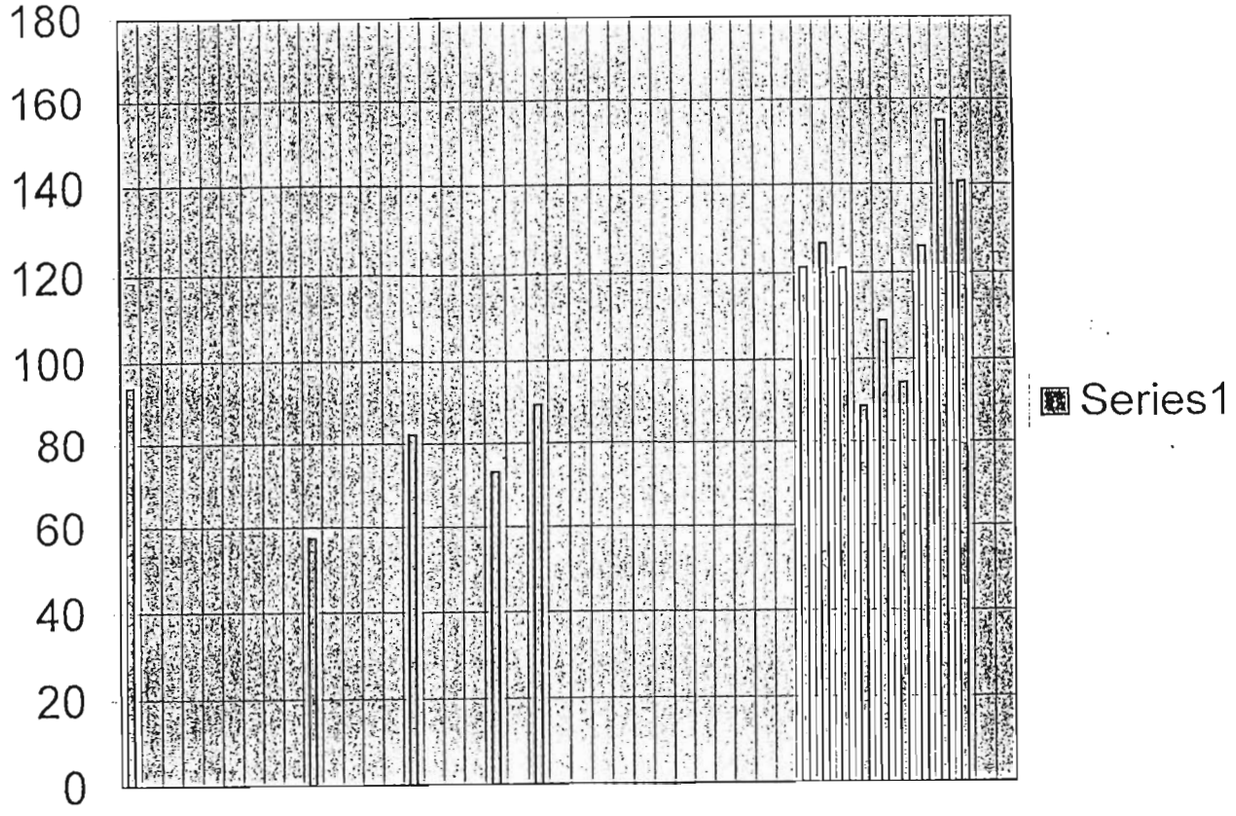
• - denotes extraneous peaks



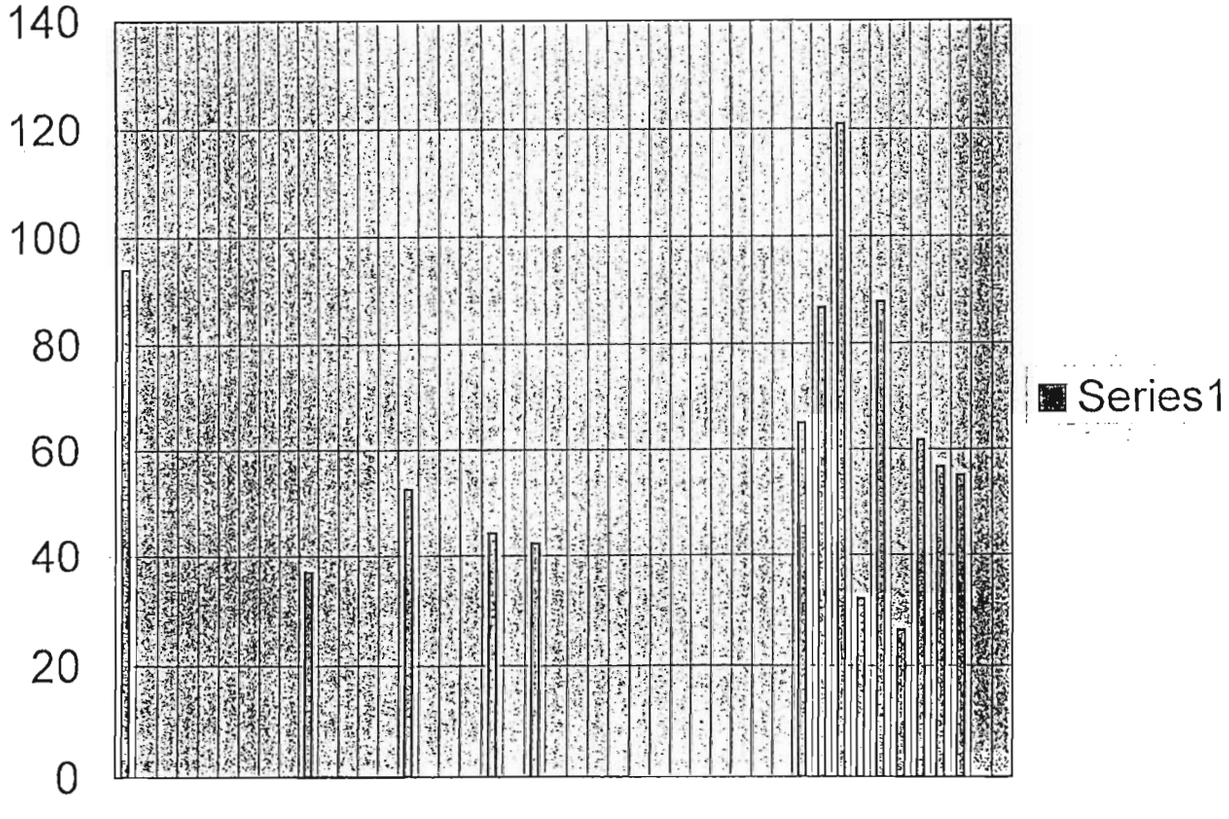
COMPOUND XXXX



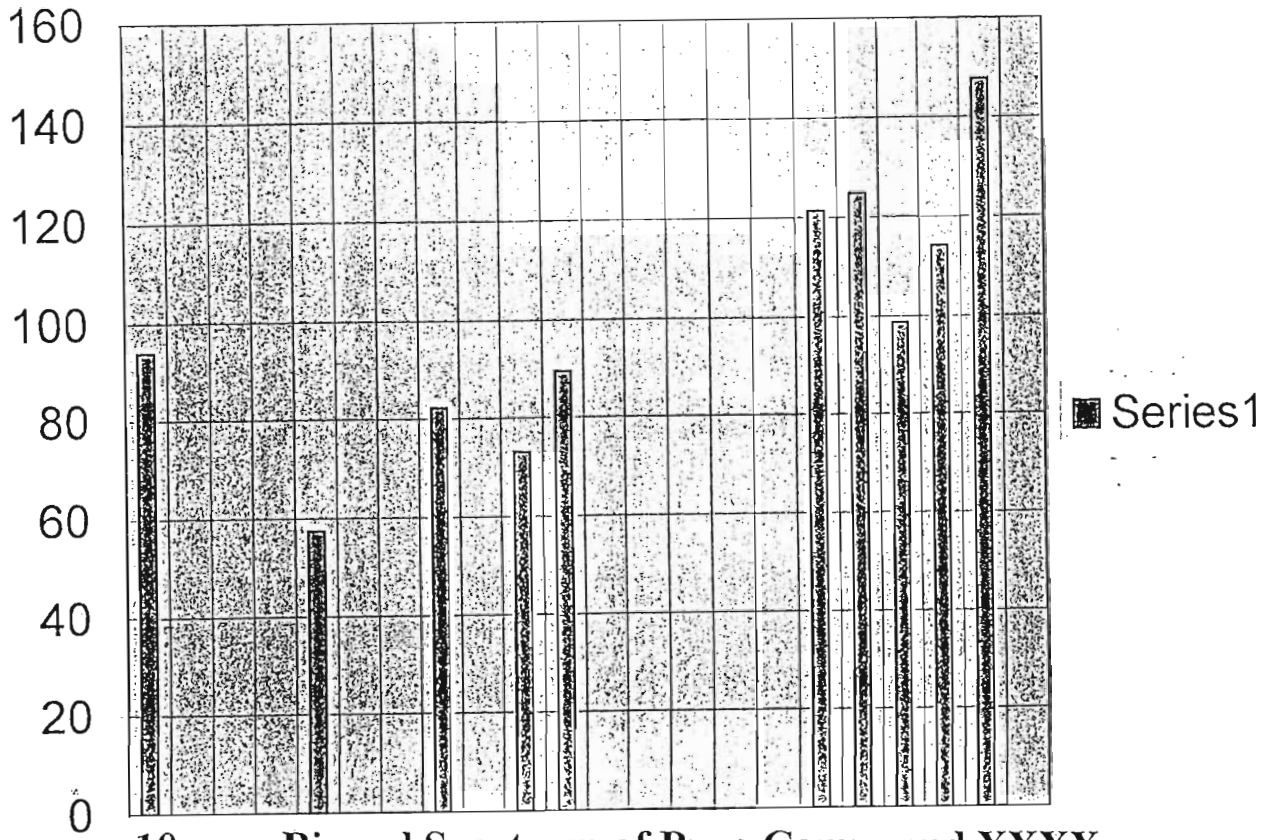
^{13}C NMR Spectrum of Compound XXXX



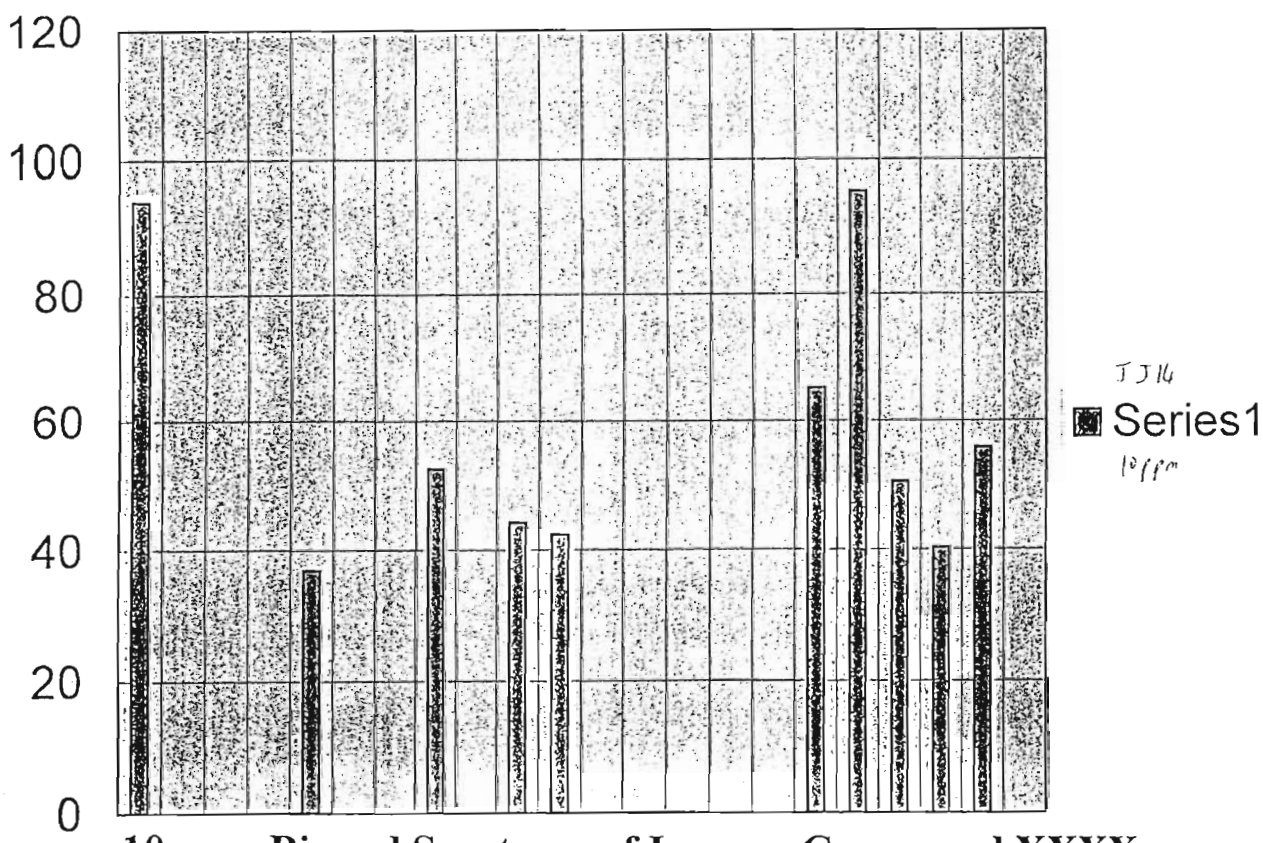
5 ppm Binned Spectrum of Pure Compound XXXX



5 ppm Binned Spectrum of Impure Compound XXXX



10 ppm Binned Spectrum of Pure Compound XXXX



10 ppm Binned Spectrum of Impure Compound XXXX

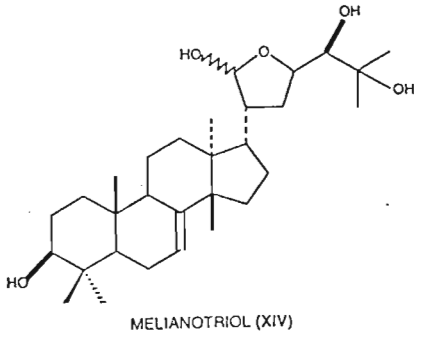
^{13}C NMR Data for Impure Compound XIV

ppm	Intensity
145.941	57.959
• 145.871	15.357
118.46	54.582
• 102.472	12.891
97.05	51.267
79.431	65.31
78.903	53.774
• 77.912	39.465
• 77.48	26.657
• 77.273	39.465
• 78.873	14.299
• 76.634	39.465
• 76.087	12.769
• 75.181	55.95
73.963	69.515
• 73.444	17.446
• 51.108	13.976
50.861	123.24
• 50.445	12.849

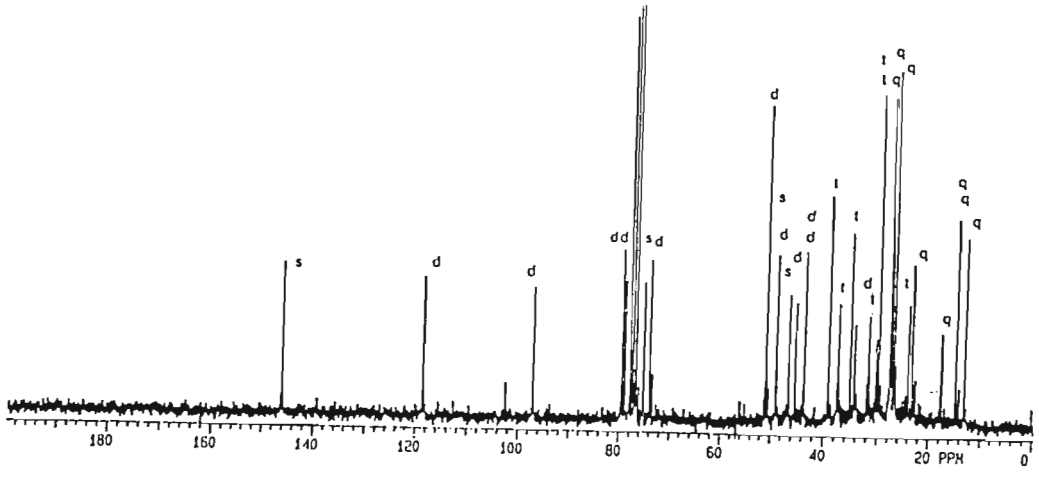
ppm	Intensity
49.058	13.934
48.928	66.511
46.502	53.248
45.372	48.532
• 43.774	14.419
43.639	66.275
43.564	13.013
39.053	89.131
• 37.486	14.758
37.304	47.012
35.098	77.182
34.284	38.757
32.009	15.334
31.647	42.028
• 30.339	33.352
• 30.24	17.638
• 29.779	137.426
• 29.439	15.797
• 27.864	12.842

ppm	Intensity
27.678	127.345
27.391	80.147
• 27.328	47.007
• 27.215	29.18
26.804	137.927
• 26.608	22.431
• 26.509	20.518
24.02	46.956
23.282	68.65
• 22.749	15.578
• 22.565	17.112
17.574	35.881
14.772	82.306
14.176	14.625
13.088	75.718

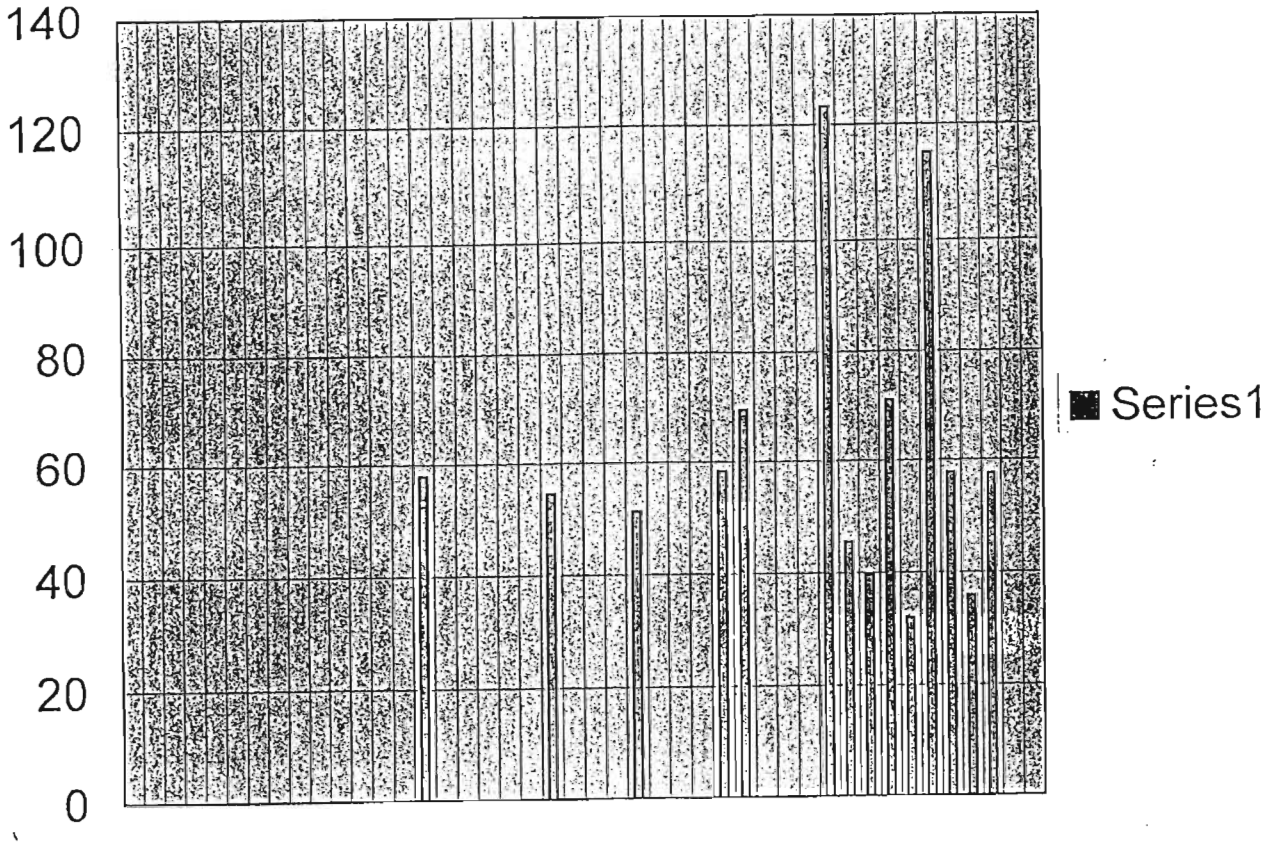
• - denotes extraneous peaks



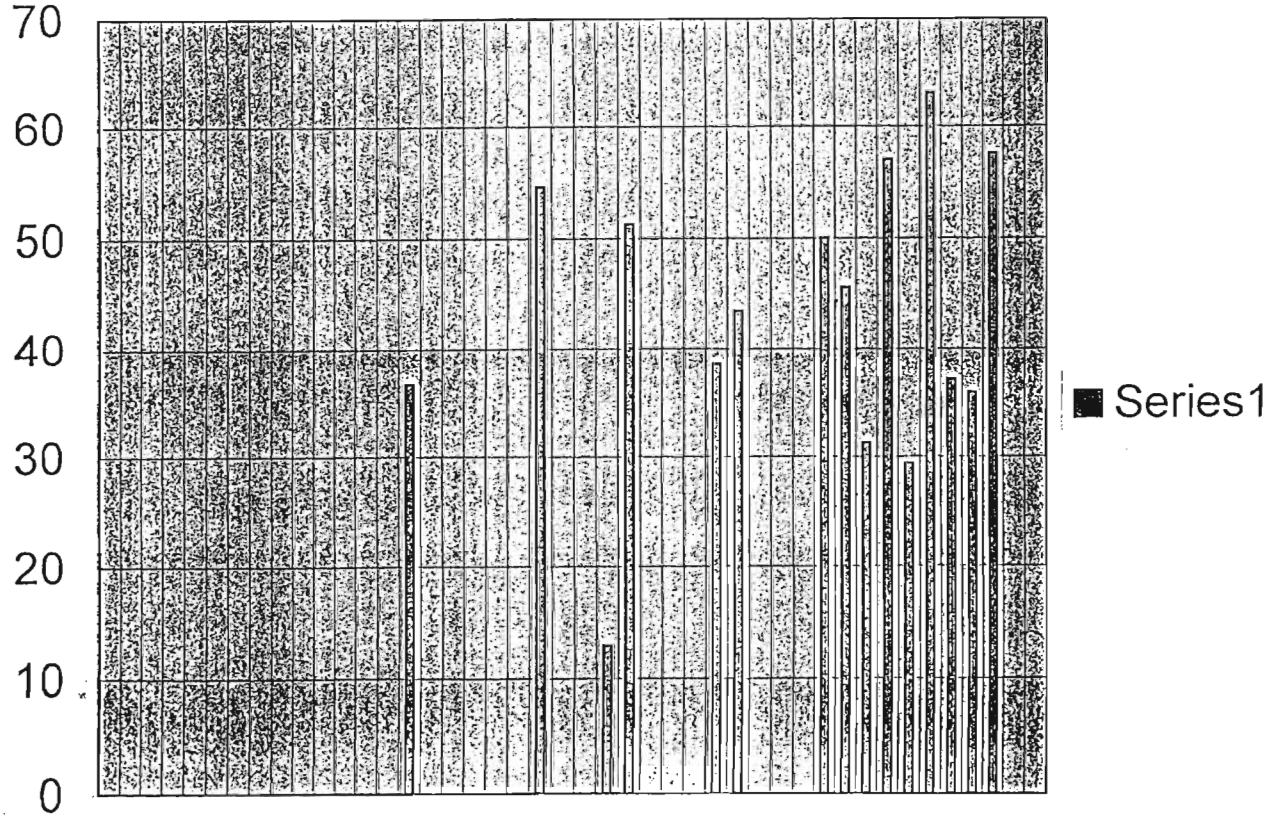
MELIANOTRIOL (XIV)



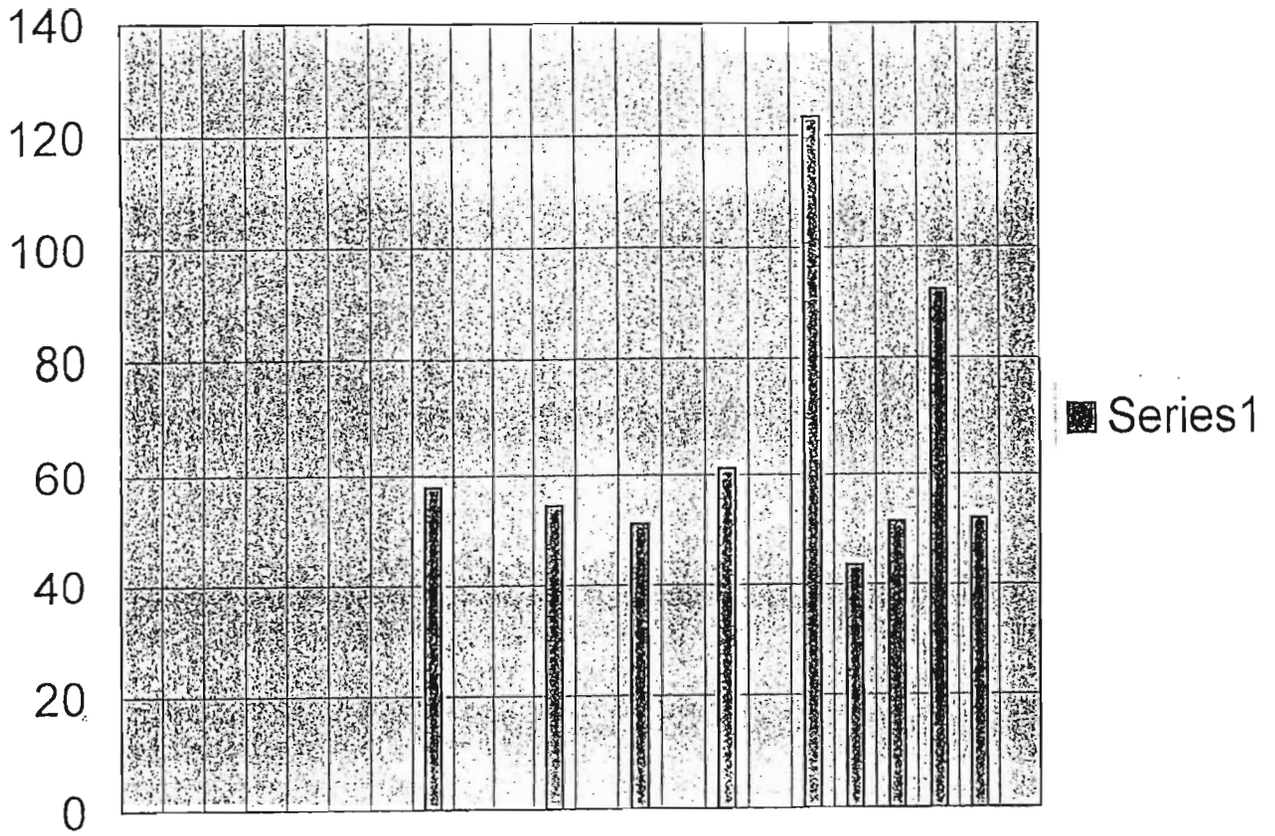
^{13}C NMR Spectrum of Compound XIV



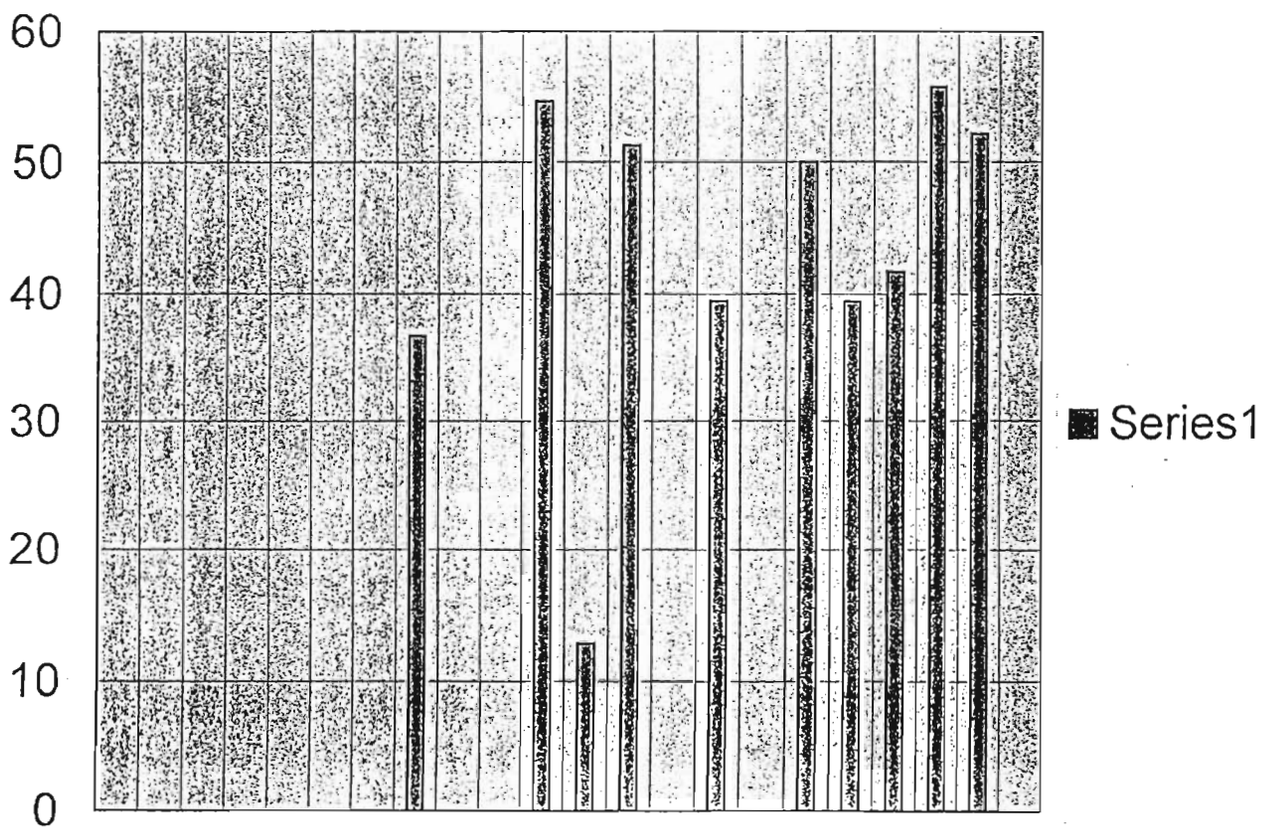
5 ppm Binned Spectrum of Pure Compound XIV



5 ppm Binned Spectrum of Impure Compound XIV



10 ppm Binned Spectrum of Pure Compound XIV



10 ppm Binned Spectrum of Impure Compound XIV

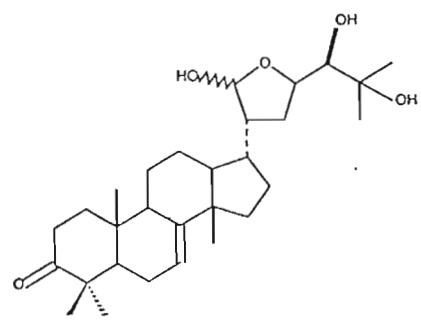
^{13}C NMR Data for Impure Compound XII

ppm	Intensity
217.607	59.901
145.777	70.429
118.071	46.15
• 116.979	32.346
96.566	45.946
• 78.466	49.142
77.845	94.708
• 77.404	61.6396
• 77.204	61.6396
• 76.565	61.6396
75.079	47.516
73.706	96.015
• 73.146	20.817
52.294	48.826
• 50.834	28.019
50.598	93.744
• 48.673	19.948
48.18	59.905
47.727	104.814

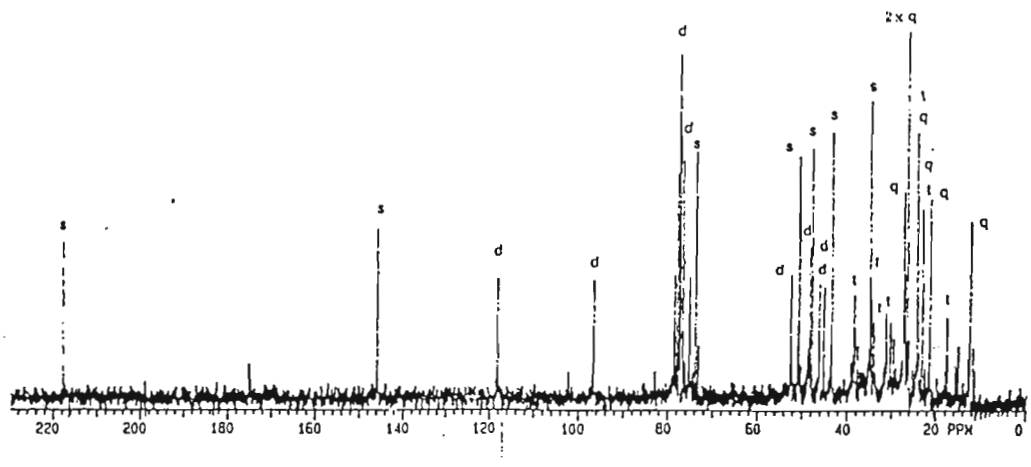
ppm	Intensity
46.135	44.089
45.068	43.534
• 43.432	27.777
43.301	103.688
38.385	39.962
• 37.639	20.804
34.877	117.763
• 34.716	43.153
34.665	38.717
34.036	30.061
31.197	32.882
30.269	29.881
• 29.475	23.907
27.277	84.722
• 27.105	48.478
• 27.038	39.563
• 26.939	20.922
• 26.63	19.77
26.452	142.159

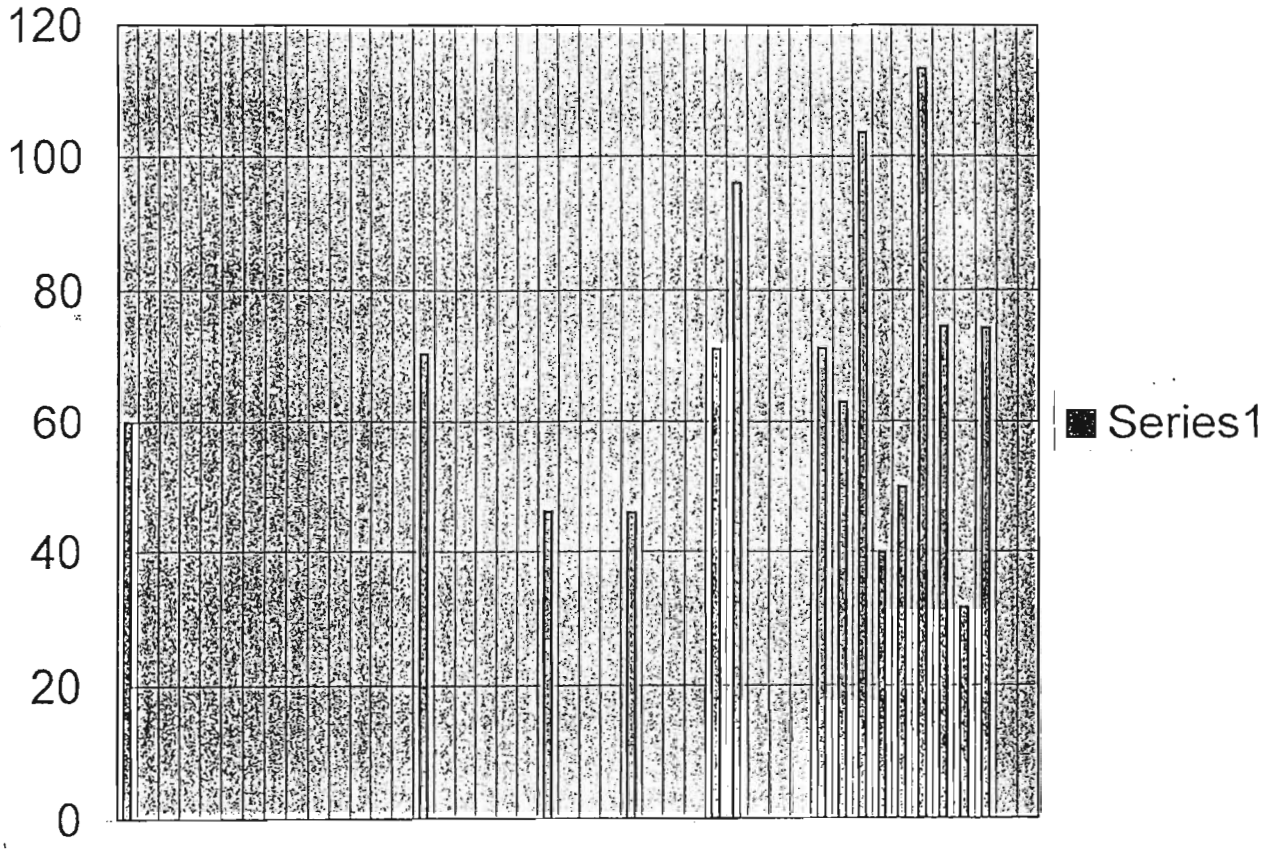
ppm	Intensity
• 26.28	23.205
24.281	102.704
24.145	37.405
23.047	73.57
21.329	84.675
17.481	31.569
• 14.949	20.64
12.479	74.283
• 11.694	20.794

• - denotes extraneous peaks

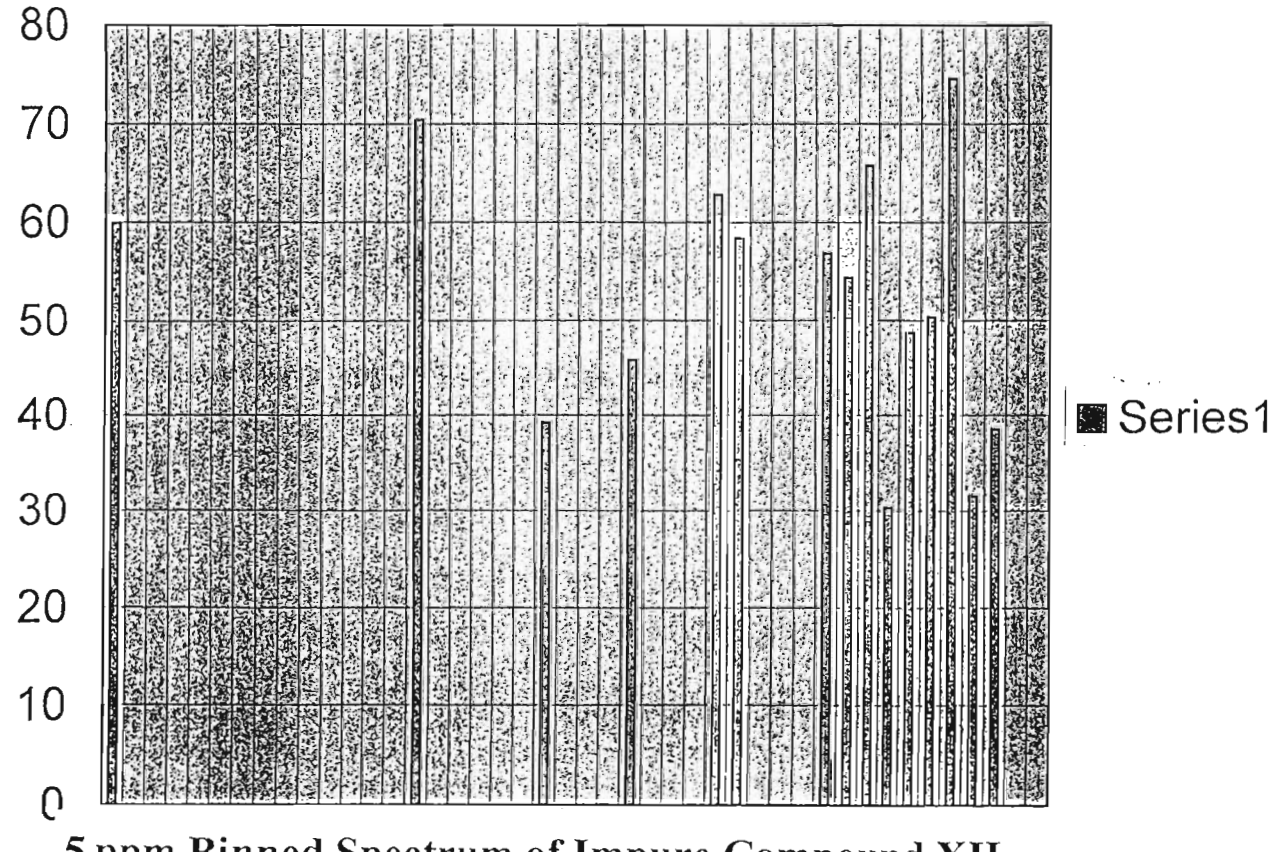


MELIANODIOL (XII)

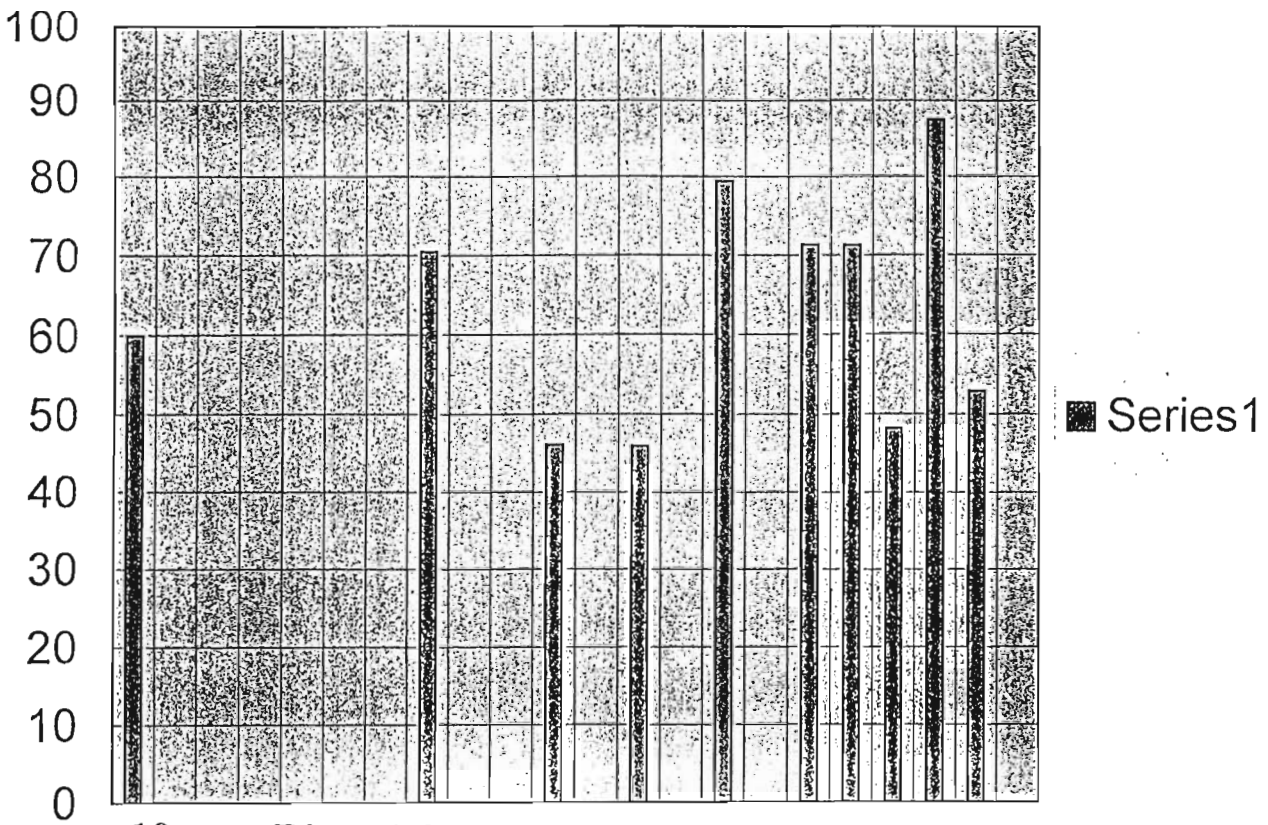




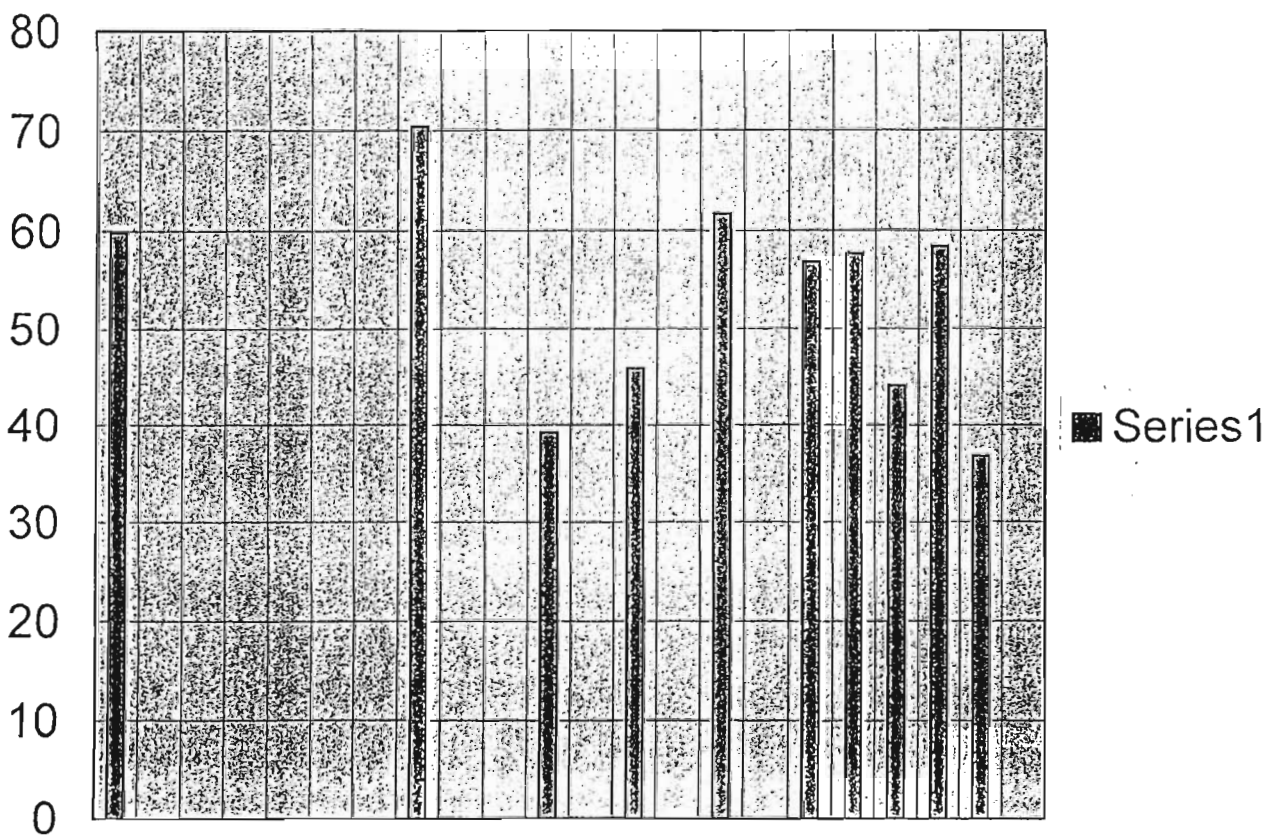
5 ppm Binned Spectrum of Pure Compound XII



5 ppm Binned Spectrum of Impure Compound XII



10 ppm Binned Spectrum of Pure Compound XII



10 ppm Binned Spectrum of Impure Compound XII

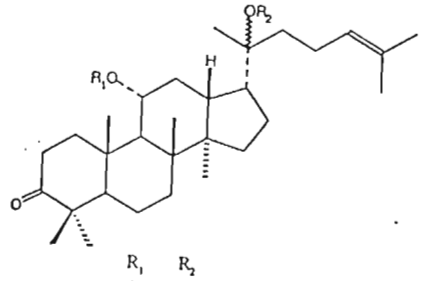
^{13}C NMR Data for Impure Compound XXXVI

ppm	Intensity
217.59	60.693
• 201.084	32.412
170.404	70.129
166.305	66.761
131.705	70.401
124.004	107.258
89.792	115.248
74.438	100.009
54.952	155.491
52.192	110.145
51.272	95.755
48.772	122.056
47.289	140.114
• 47.259	163.409
45.627	108.671
39.452	148.122
39.359	108.951
36.755	20.082

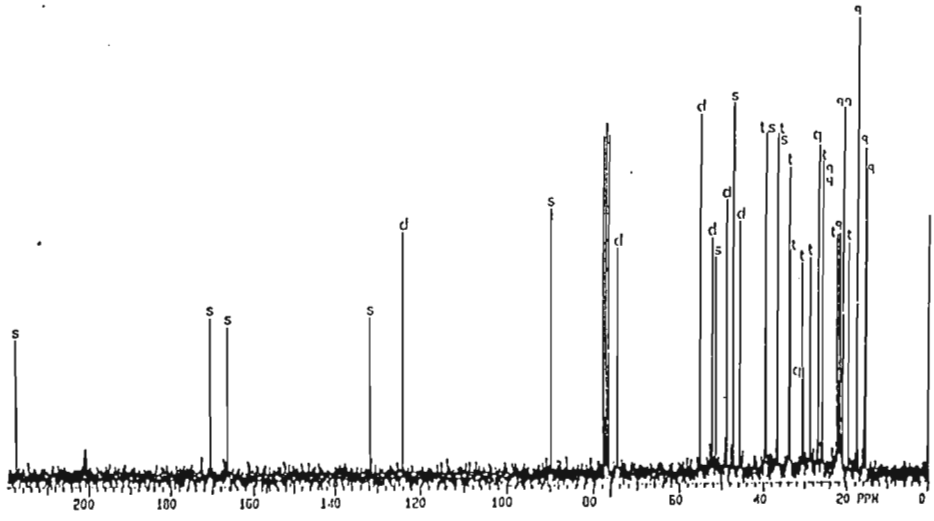
ppm	Intensity
36.668	147.818
• 36.539	77.549
33.844	131.325
33.562	95.122
30.625	90.071
30.235	44.658
28.707	92.98
• 28.625	17.848
26.779	142.897
25.766	103.39
• 25.694	133.751
22.427	102.654
21.829	103.656
21.672	94.938
20.98	157.406
19.58	98.729
17.596	195.809
• 16.031	18.129

ppm	Intensity
15.739	139.777
15.259	130.294
• 0.001	110.86

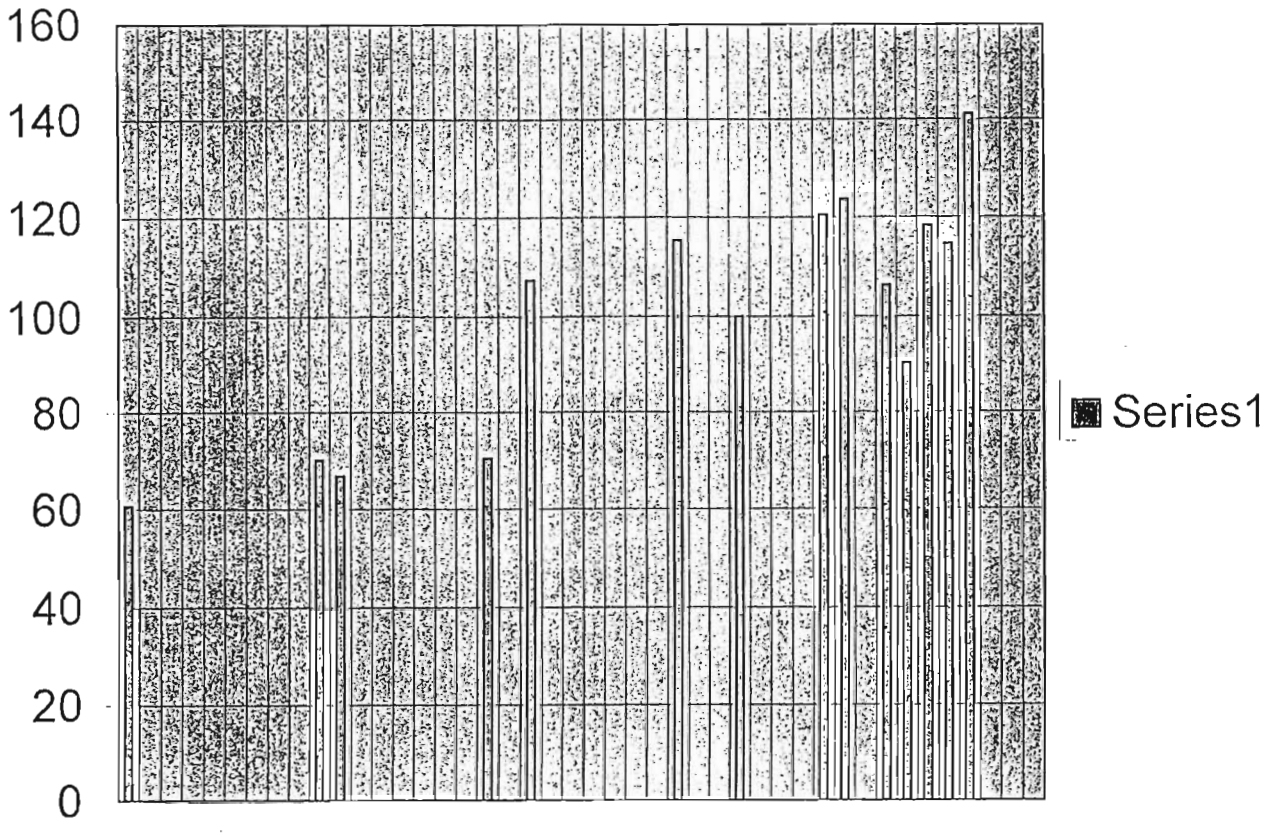
• - denotes extraneous peaks



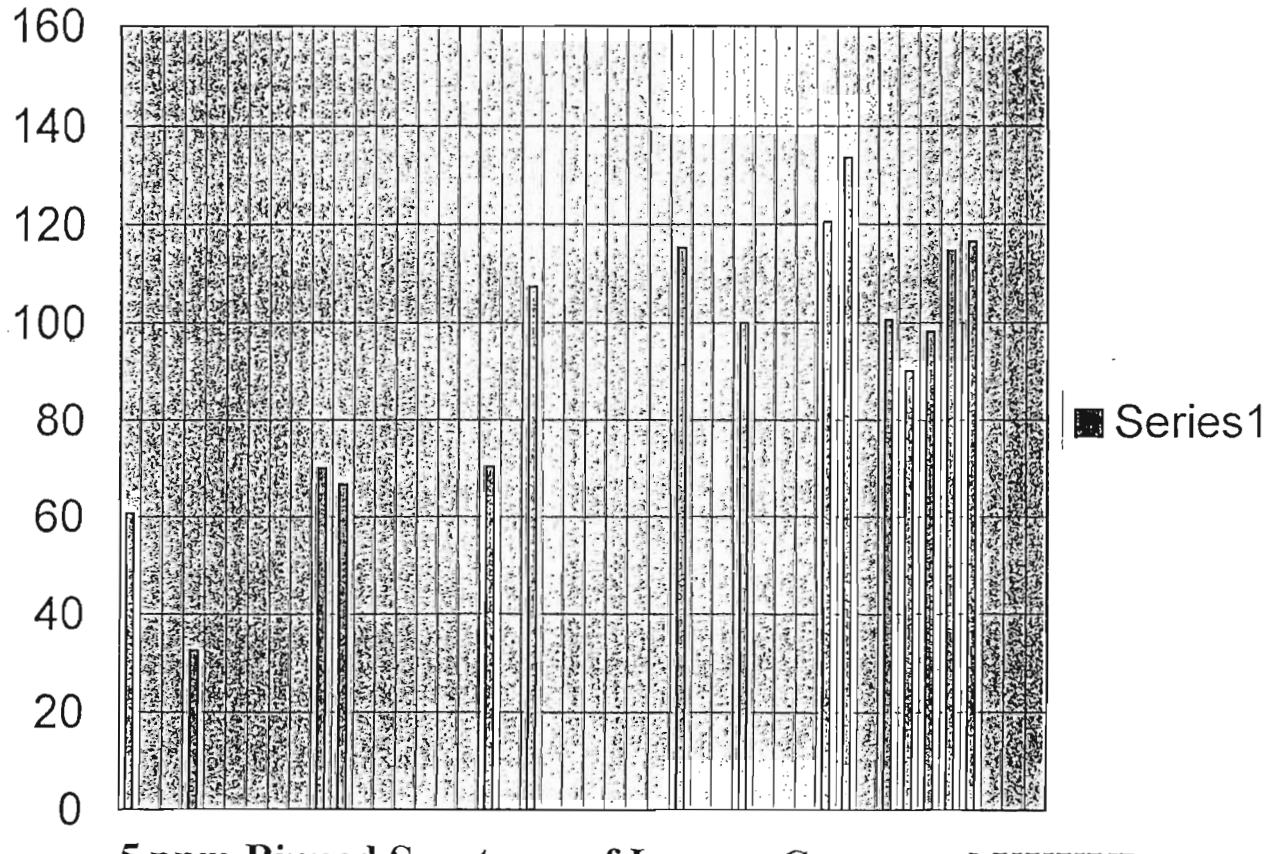
R_1 R_2
Ac Ac
COMPOUND XXXVI



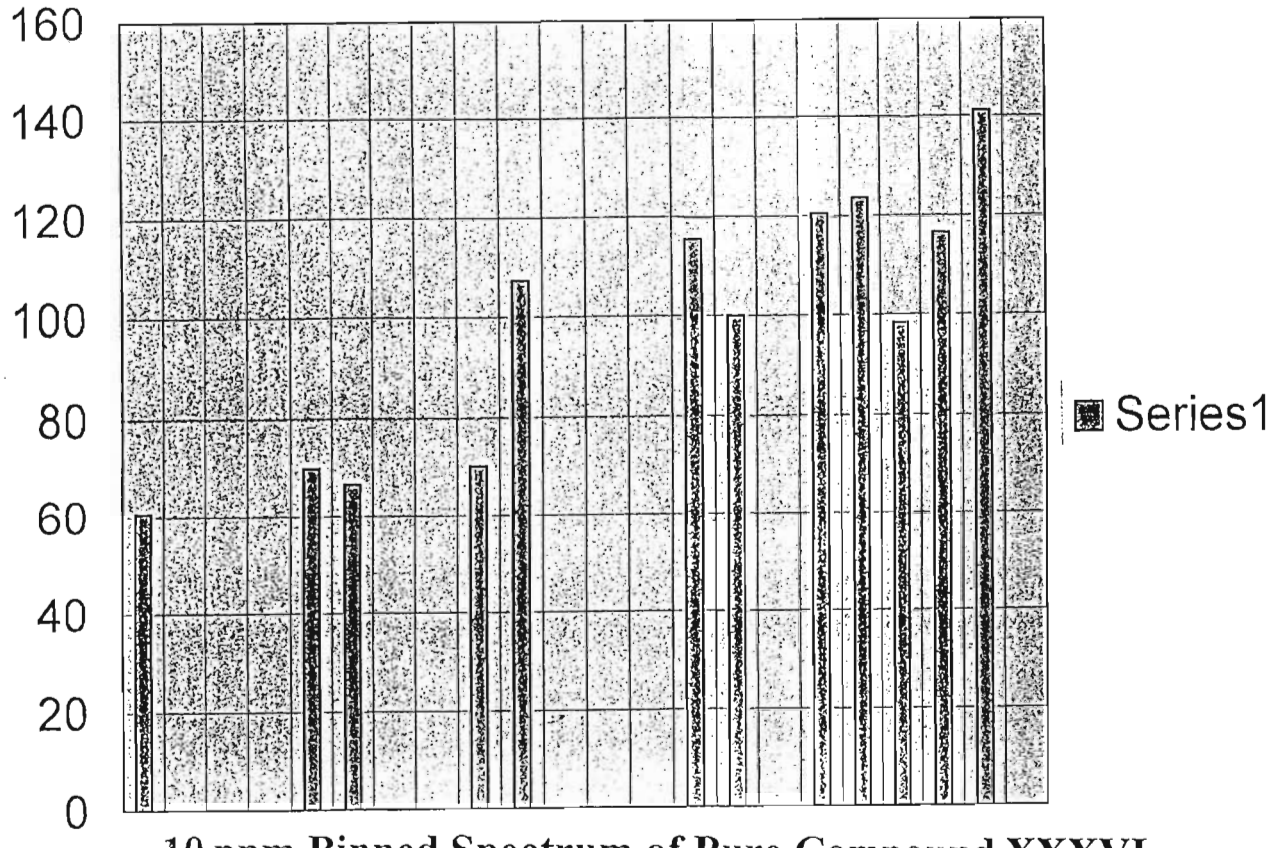
^{13}C NMR Spectrum of Compound XXXVI



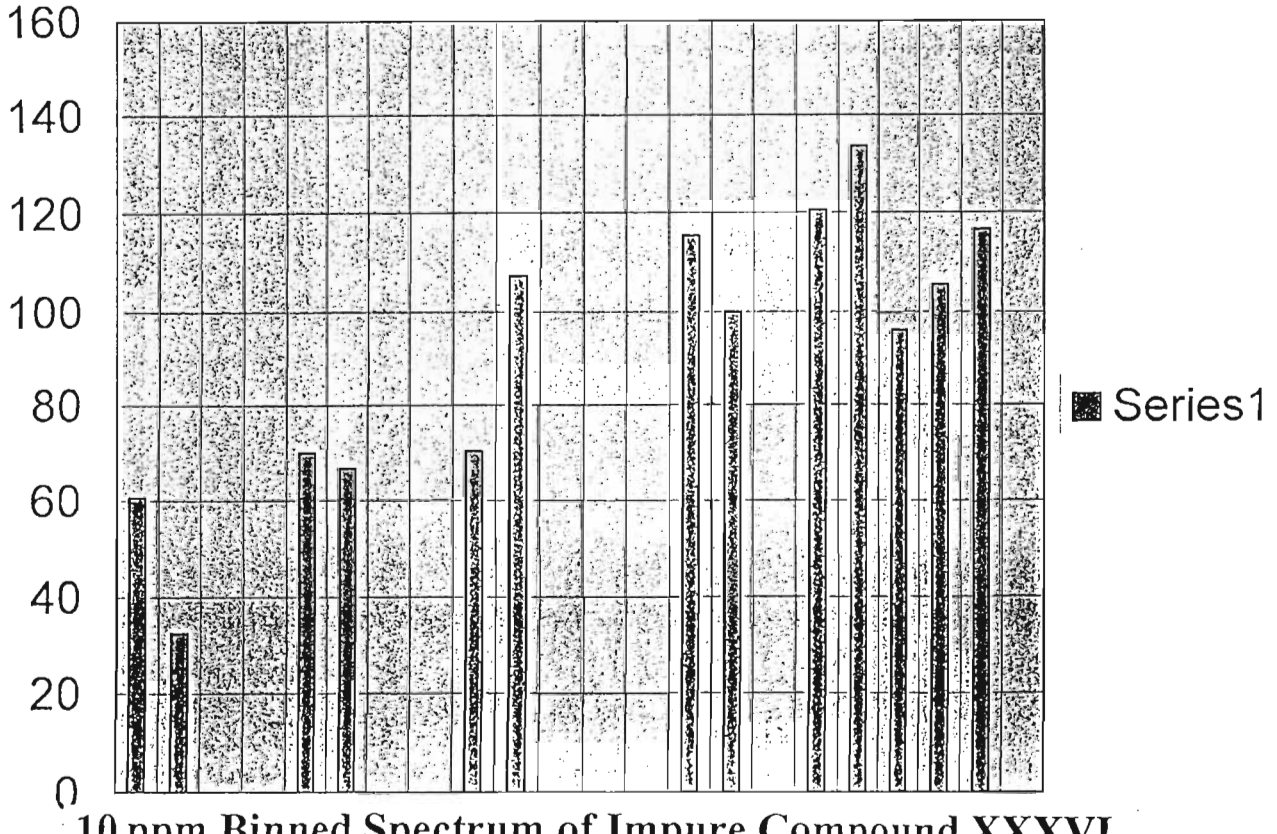
5 ppm Binned Spectrum of Pure Compound XXXVI



5 ppm Binned Spectrum of Impure Compound XXXVI



10 ppm Binned Spectrum of Pure Compound XXXVI



10 ppm Binned Spectrum of Impure Compound XXXVI

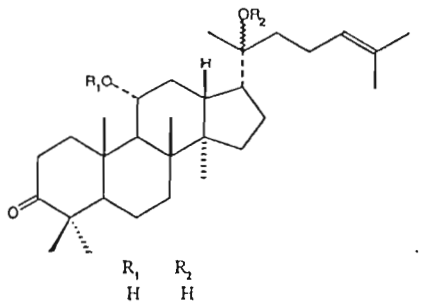
^{13}C NMR Data for Impure Compound XXXIV

ppm	Intensity
217.835	27.595
132.007	31.501
• 130.857	18.908
• 128.769	19.293
124.784	90.619
• 77.2	12.847
74.641	54.246
70.663	83.03
• 68.138	15.767
55.272	89.141
53.314	96.657
51.572	55.078
49.331	98.825
47.942	92.135
47.347	65.763
39.712	85.78
39.624	68.493

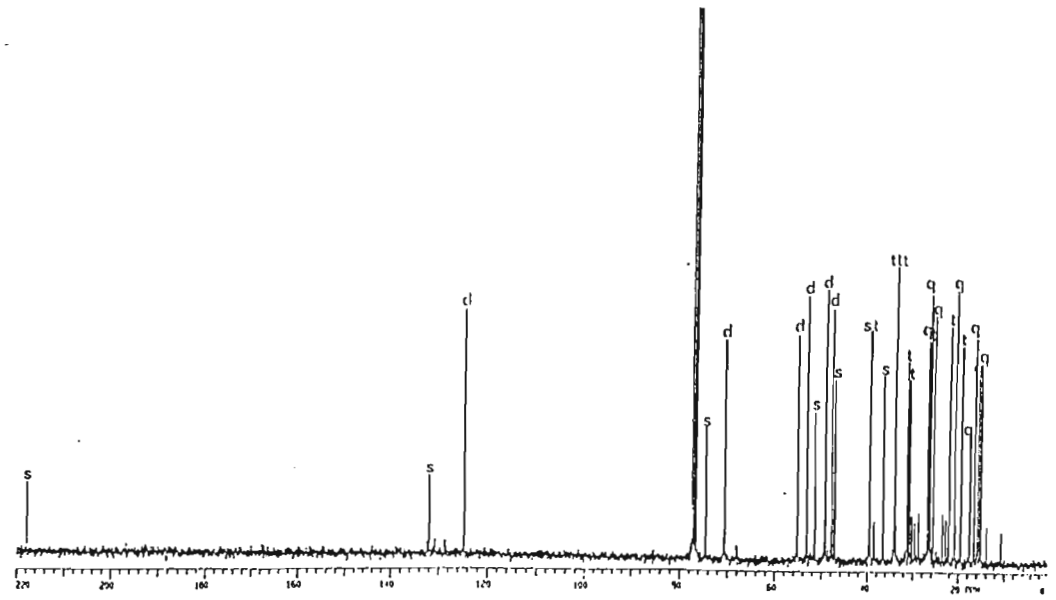
ppm	Intensity
• 38.705	14.757
36.77	67.373
34.294	78.026
34.086	90.16
34.04	108.507
31.456	75.194
30.932	72.249
• 30.331	16.211
• 29.677	13.811
• 29.897	16.995
27.035	80.175
26.682	97.177
26.423	80.589
25.751	89.552
• 23.721	16.68
• 22.958	14.771
22.333	85.657

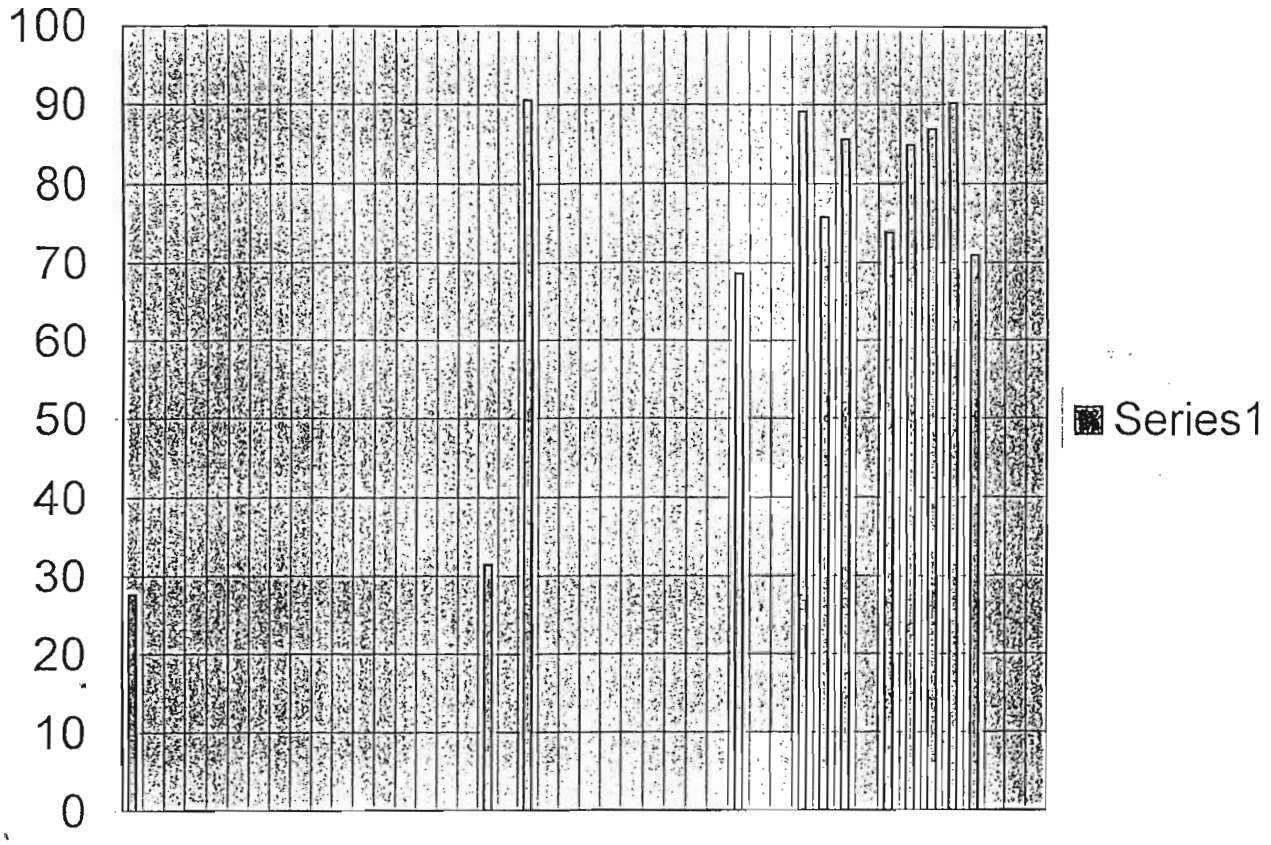
ppm	Intensity
22.333	85.657
21.003	99.247
19.63	78.924
17.752	44.917
16.732	81.069
15.916	77.606
15.404	72.053
• 14.03	12.117

• - denotes extraneous peaks

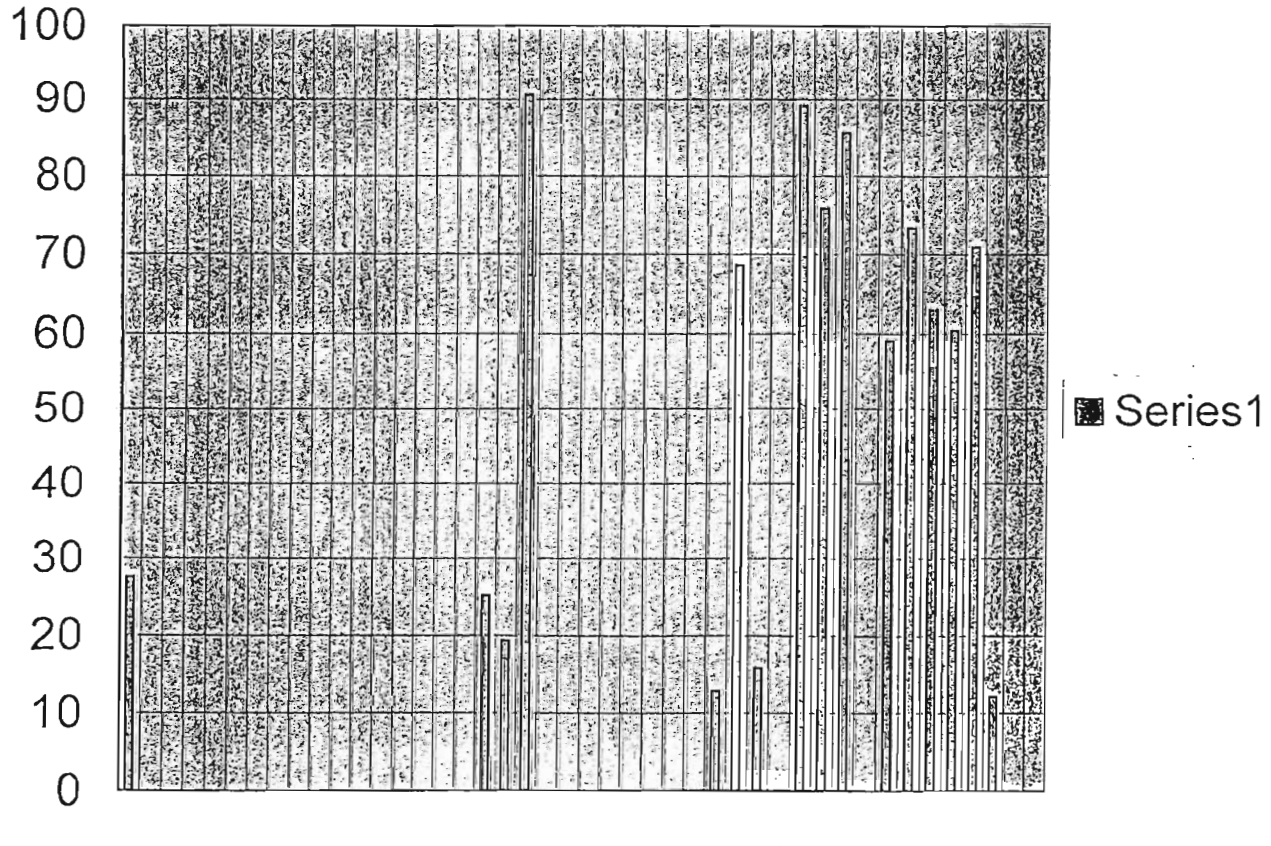


COMPOUND XXXIV

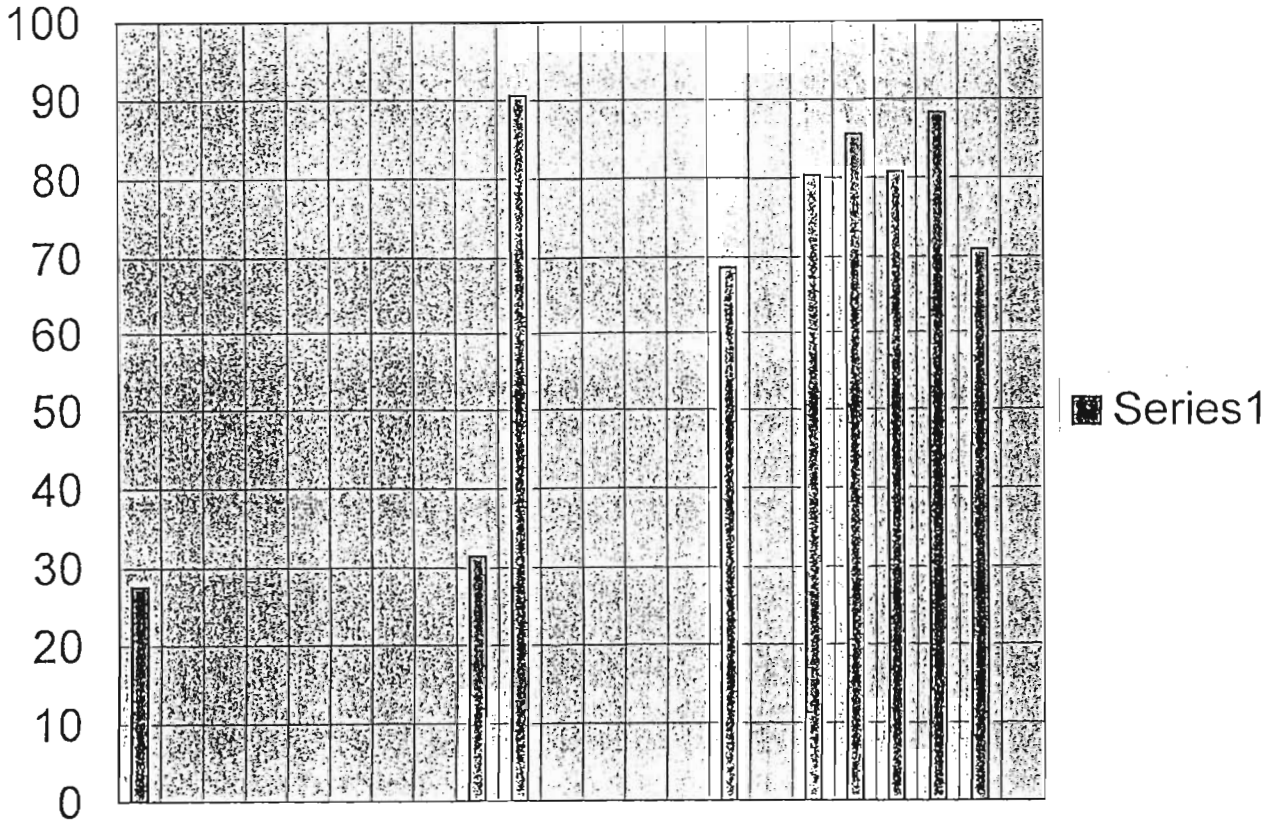




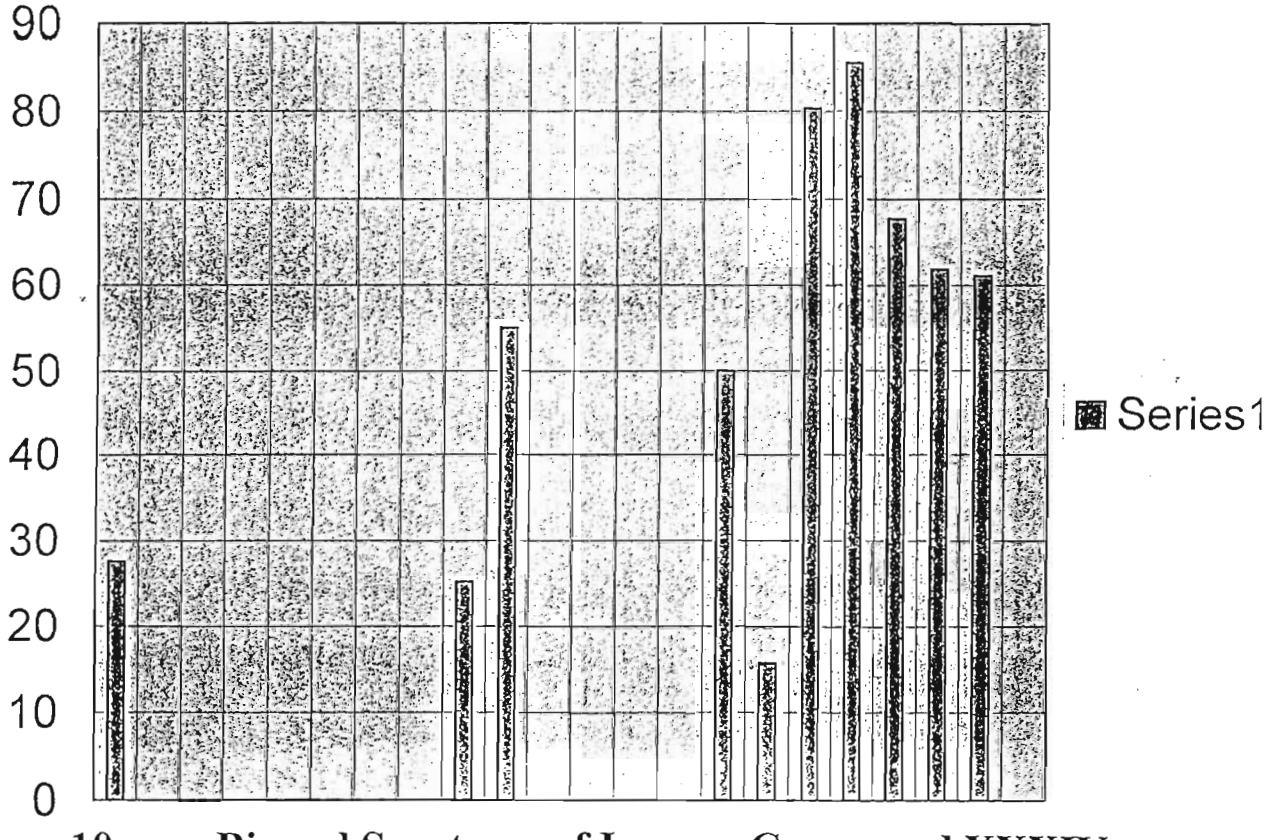
5 ppm Binned Spectrum of Pure Compound XXXIV



5 ppm Binned Spectrum of Impure Compound XXXIV



10 ppm Binned Spectrum of Pure Compound XXXIV



10 ppm Binned Spectrum of Impure Compound XXXIV

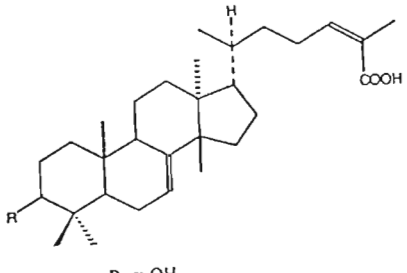
^{13}C NMR Data for Impure Compound XXXXIII

ppm	Intensity
169.202	11.468
146.493	18.733
143.969	29.112
127.279	10.529
118.541	26.624
75.701	29.228
• 55.196	22.074
• 54.653	46.033
• 54.107	69.664
• 53.563	47.927
53.319	42.174
• 53.017	23.783
51.646	19.783
49.118	35.99
44.812	37.829
43.91	24.48
37.73	23.945

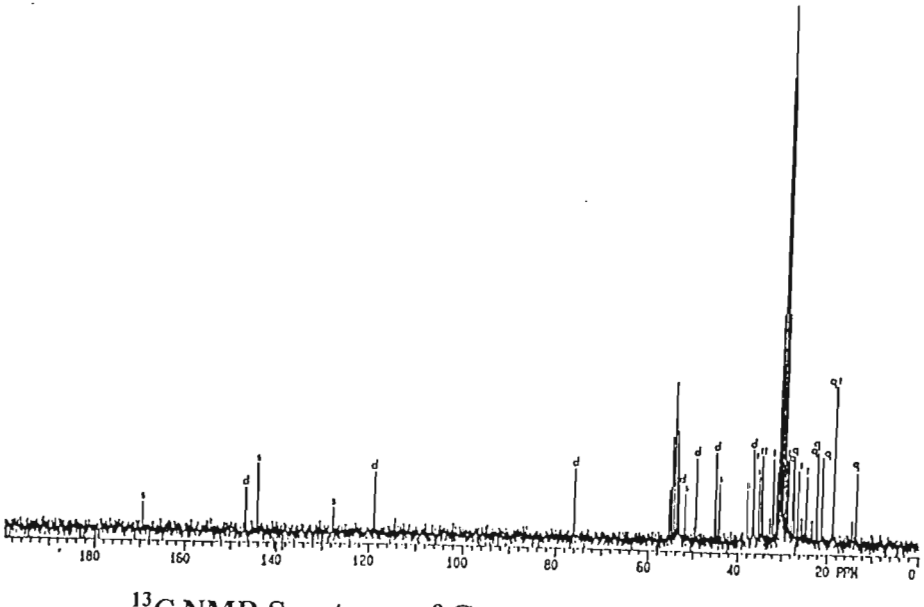
ppm	Intensity
36.532	40.052
36.125	35.179
35.086	28.079
34.413	37.272
34.313	35.947
• 32.328	10.435
31.702	35.879
• 30.96	31.848
• 30.573	98.137
• 30.189	202.79
• 30.091	76.11
• 29.927	19.133
• 29.803	238.604
• 29.664	12.113
• 29.552	17.146
• 29.418	202.308
• 29.034	99.601

ppm	Intensity
28.647	36.455
• 28.535	34.789
28.208	34.852
27.453	33.78
26.95	35.639
26.126	31.33
• 25.331	11.194
24.324	28.568
22.04	34.735
22.001	39.917
20.903	37.132
18.408	67.74
13.216	30.956

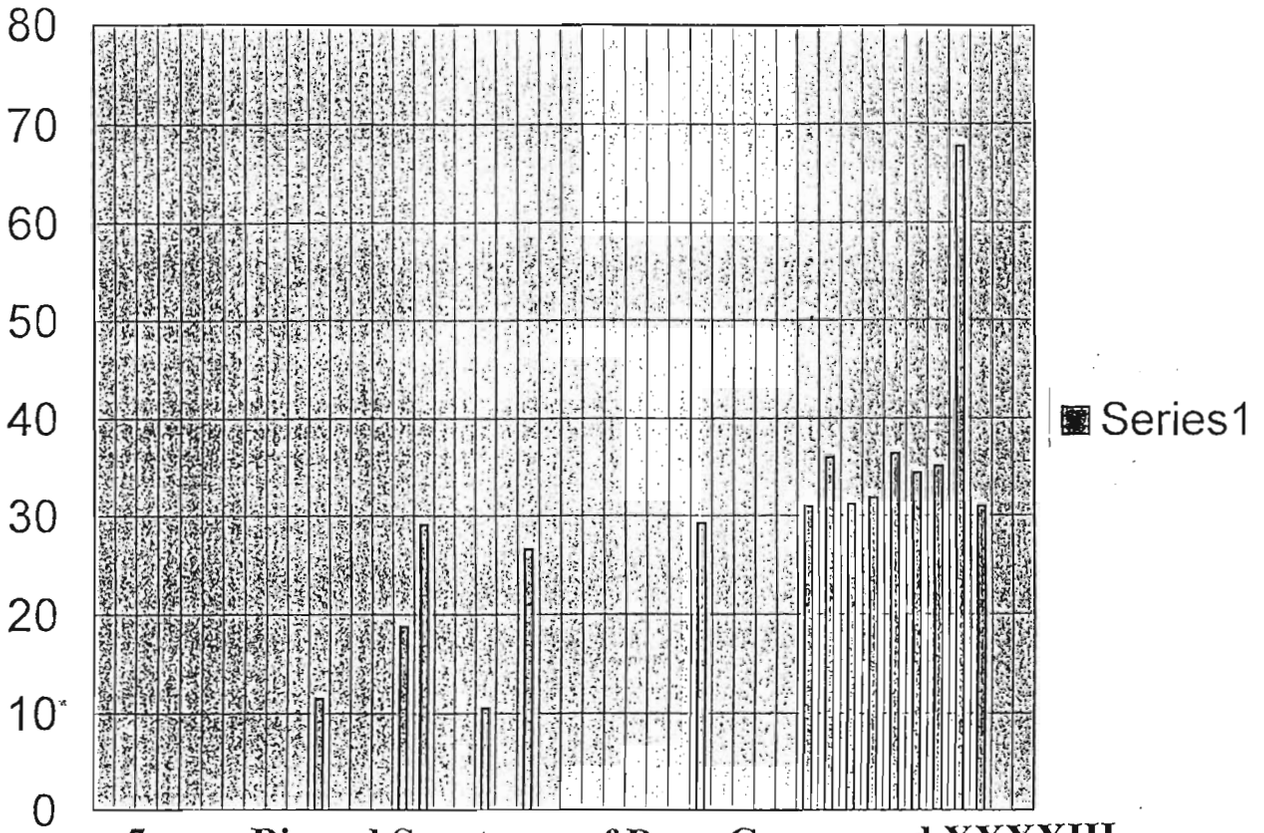
• - denotes extraneous peaks



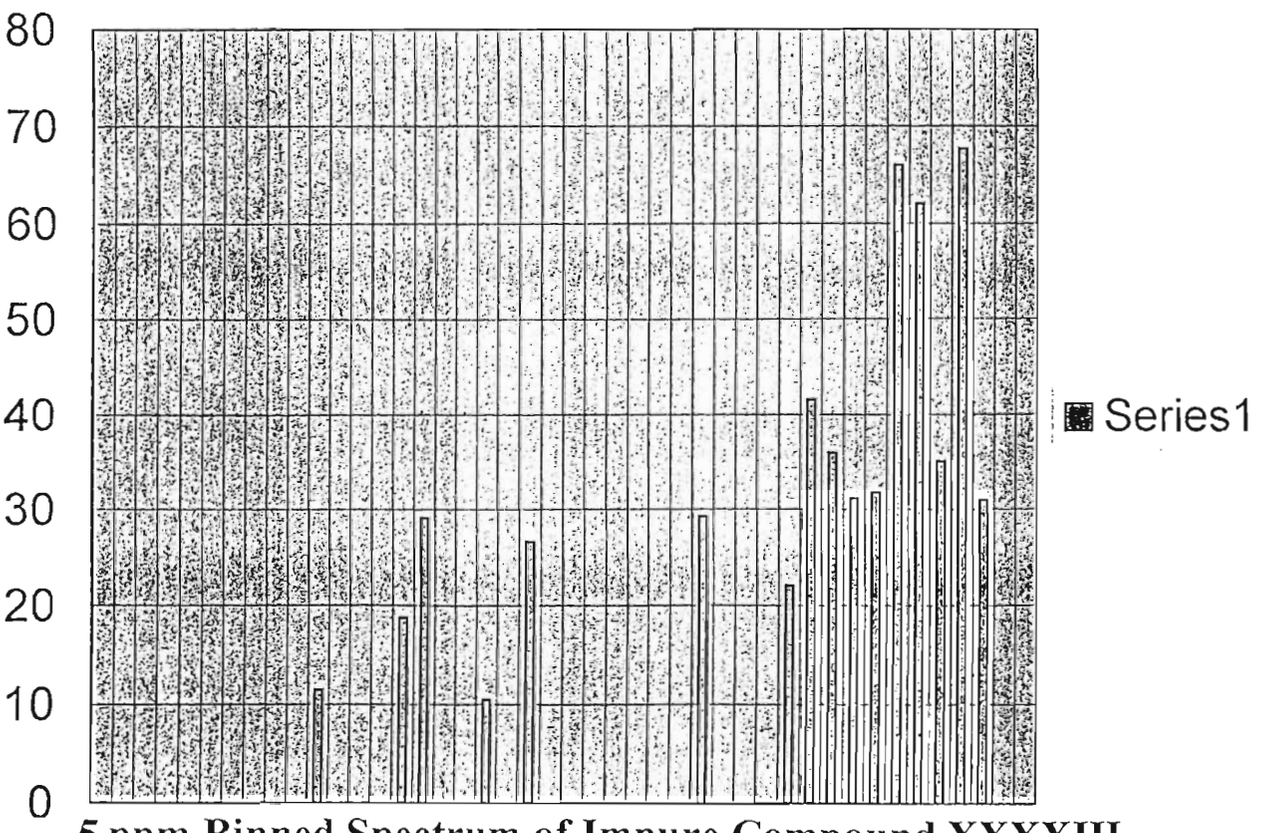
R=α-OH
COMPOUND XXXXIII



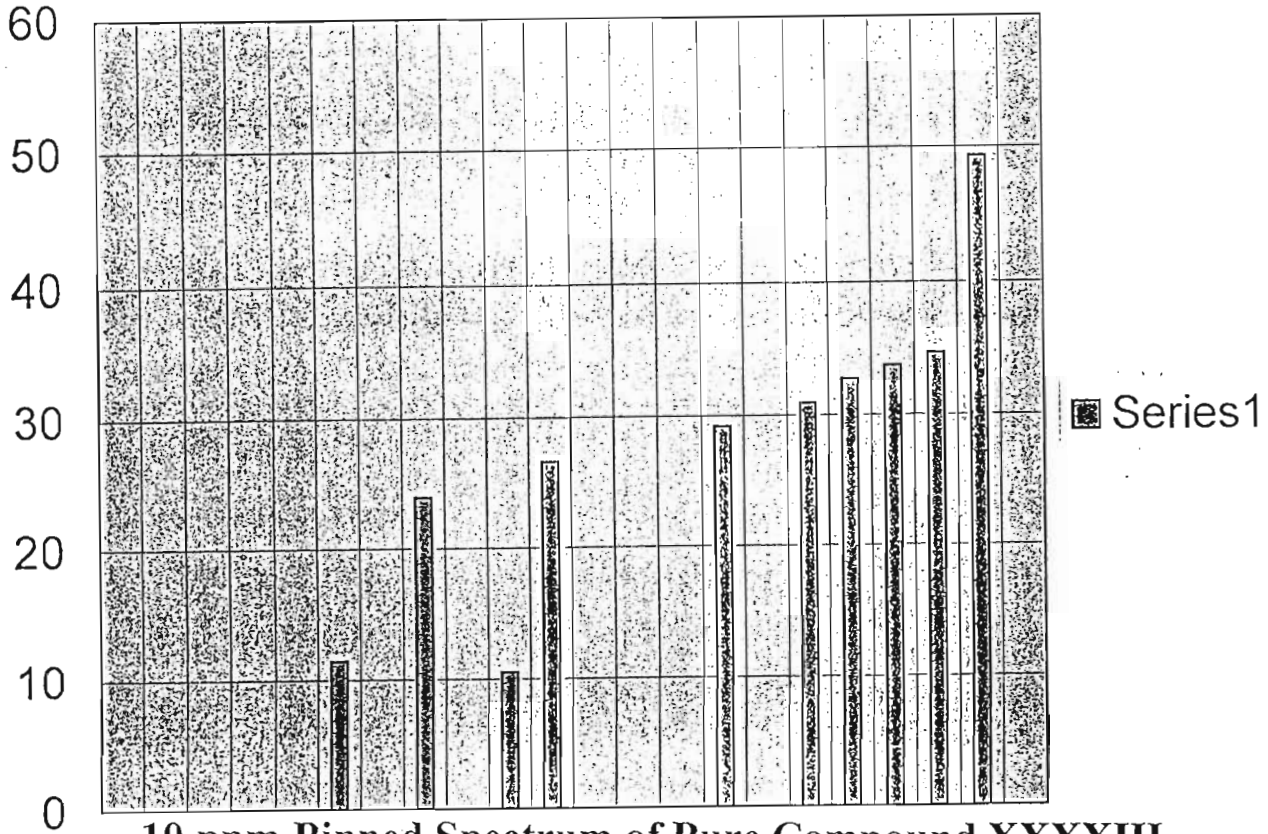
^{13}C NMR Spectrum of Compound XXXXIII



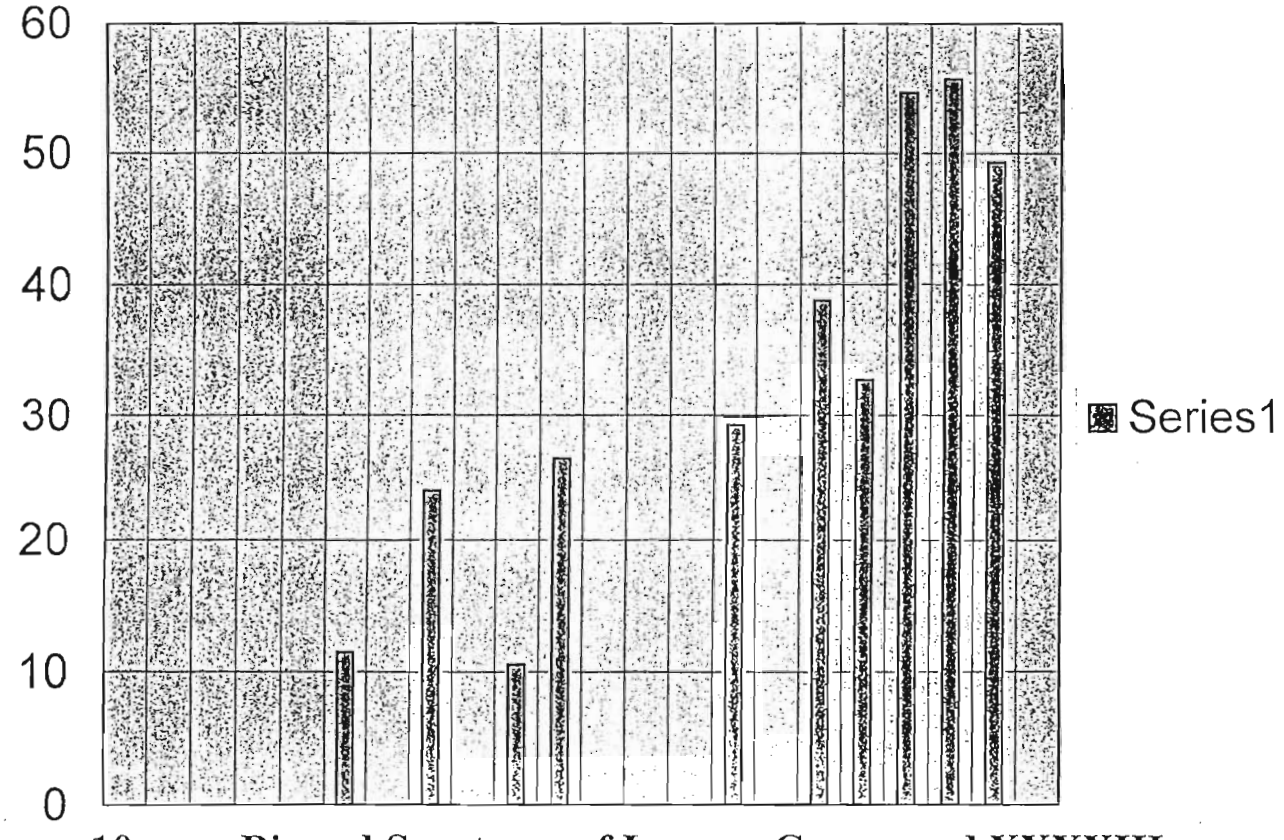
5 ppm Binned Spectrum of Pure Compound XXXXIII



5 ppm Binned Spectrum of Impure Compound XXXXIII



10 ppm Binned Spectrum of Pure Compound XXXXIII



10 ppm Binned Spectrum of Impure Compound XXXXIII

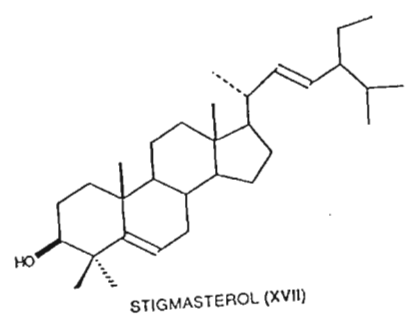
^{13}C NMR Data for Impure Compound XVII

ppm	Intensity
141.049	59.033
138.642	44.2
129.537	42.059
121.885	82.837
71.737	86.572
56.967	44.427
• 56.858	49.007
• 56.15	56.628
56.02	43.857
51.317	55.042
50.182	66.053
• 45.863	62.236
42.364	41.645
42.283	80.215
40.584	45.589
39.82	43.248

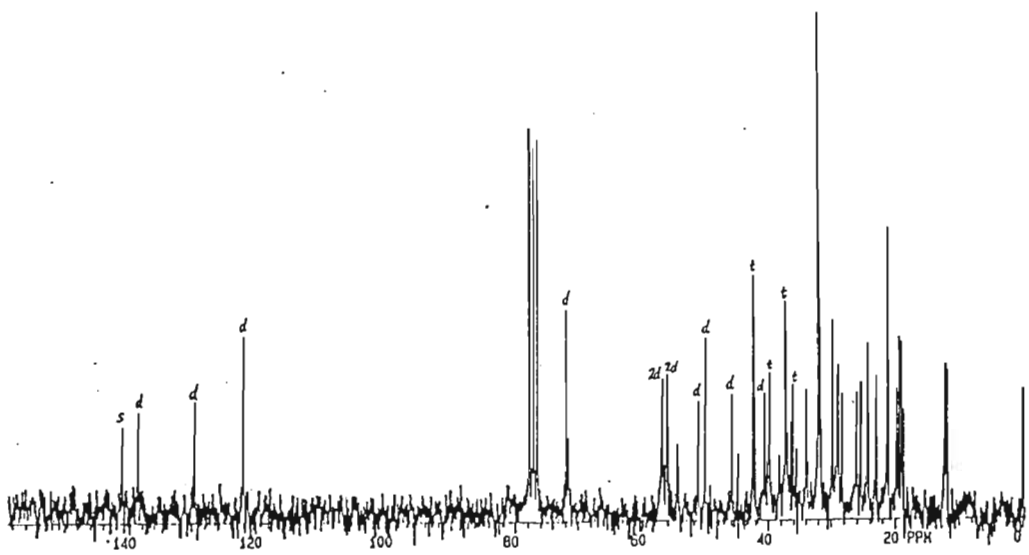
ppm	Intensity
• 39.733	38.323
37.3	75.999
36.518	54.37
• 36.195	50.22
• 33.959	33.99
31.903	205
31.575	79.18
• 29.136	62.036
28.956	42.722
• 28.284	44.738
• 26.071	46.4
25.427	49.495
• 24.372	39.847
24.308	49.185
• 23.049	61.203
21.254	59.77

ppm	Intensity
21.142	66.328
21.065	73.062
• 19.859	57.493
• 19.391	88.474
19.059	66.083
• 19.009	58.646
• 18.799	46.901
12.266	58.229
• 12.041	48.33
11.976	55.13
• 11.857	51.286

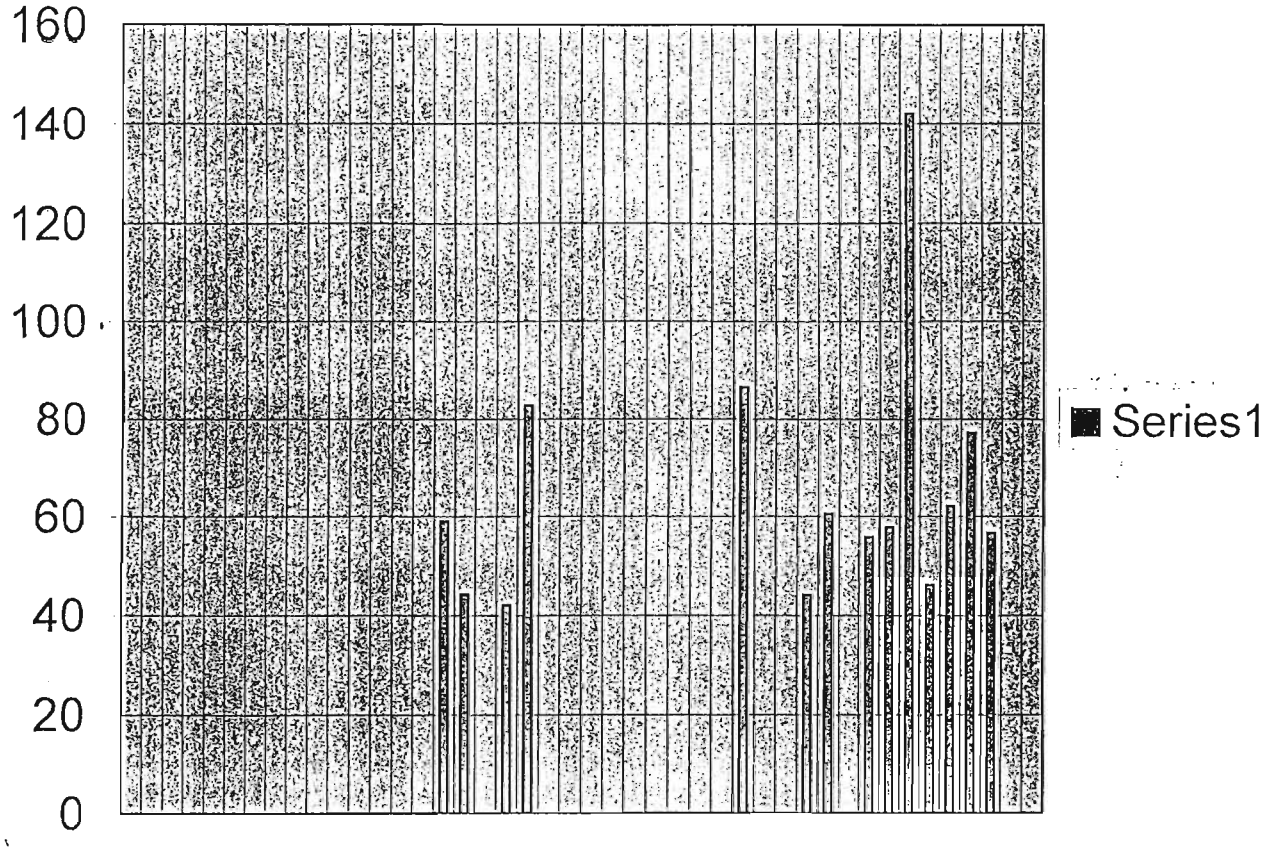
• - denotes extraneous peaks



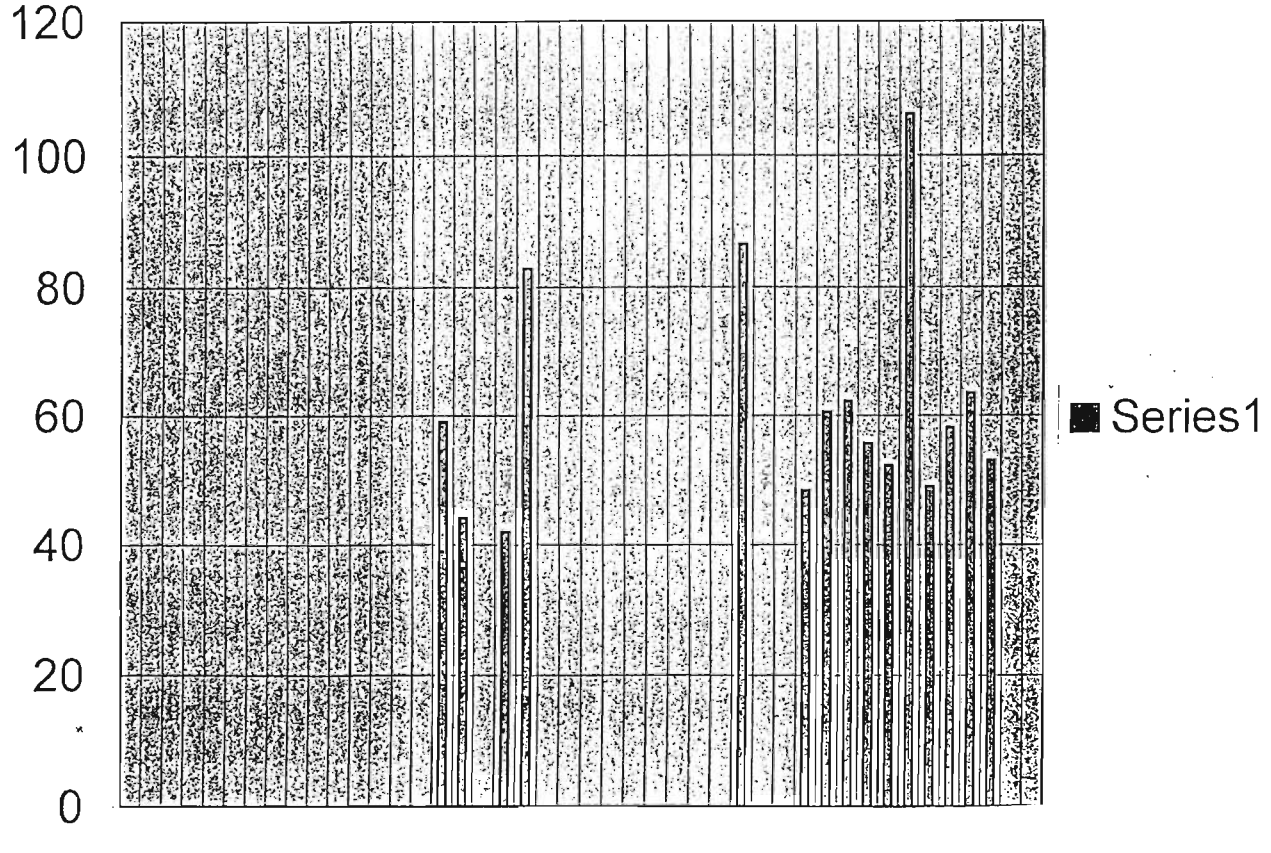
STIGMASTEROL (XVII)



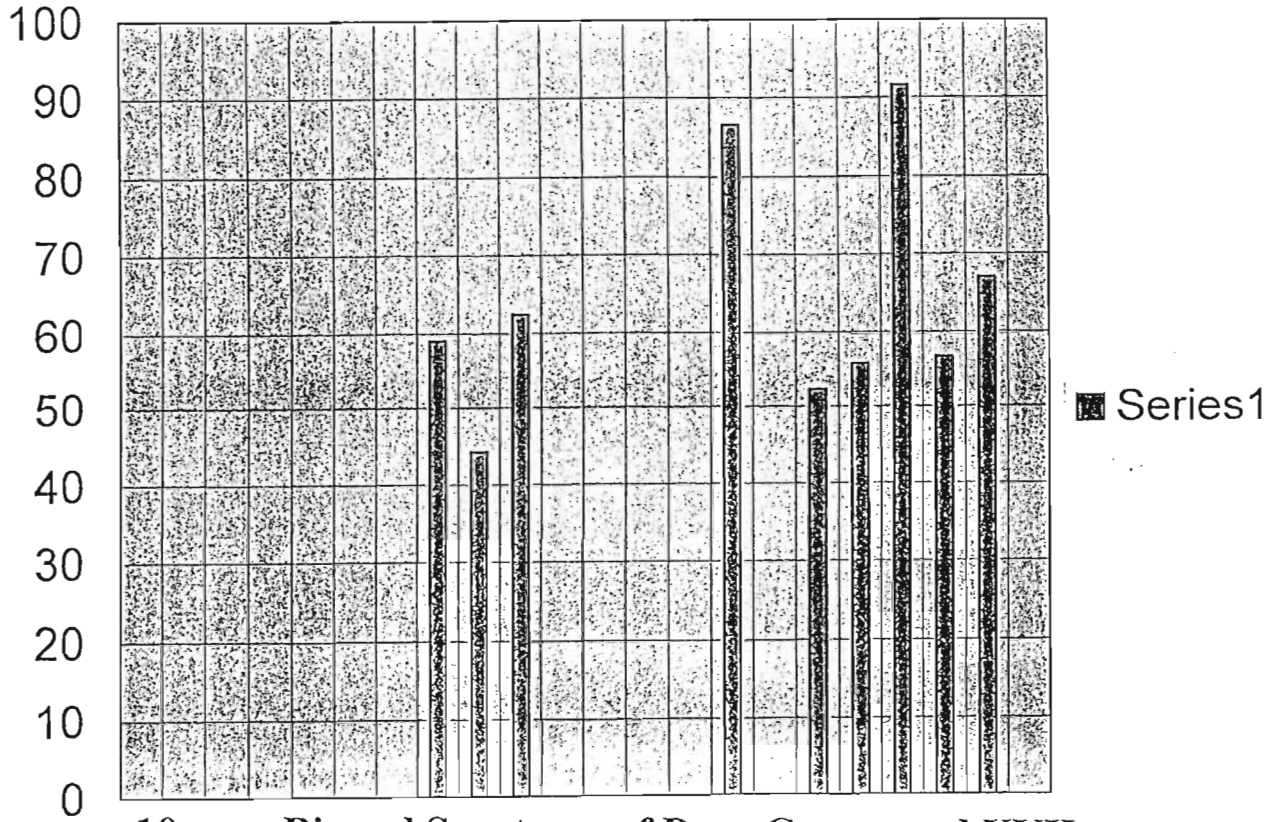
^{13}C NMR Spectrum of Compound XVII



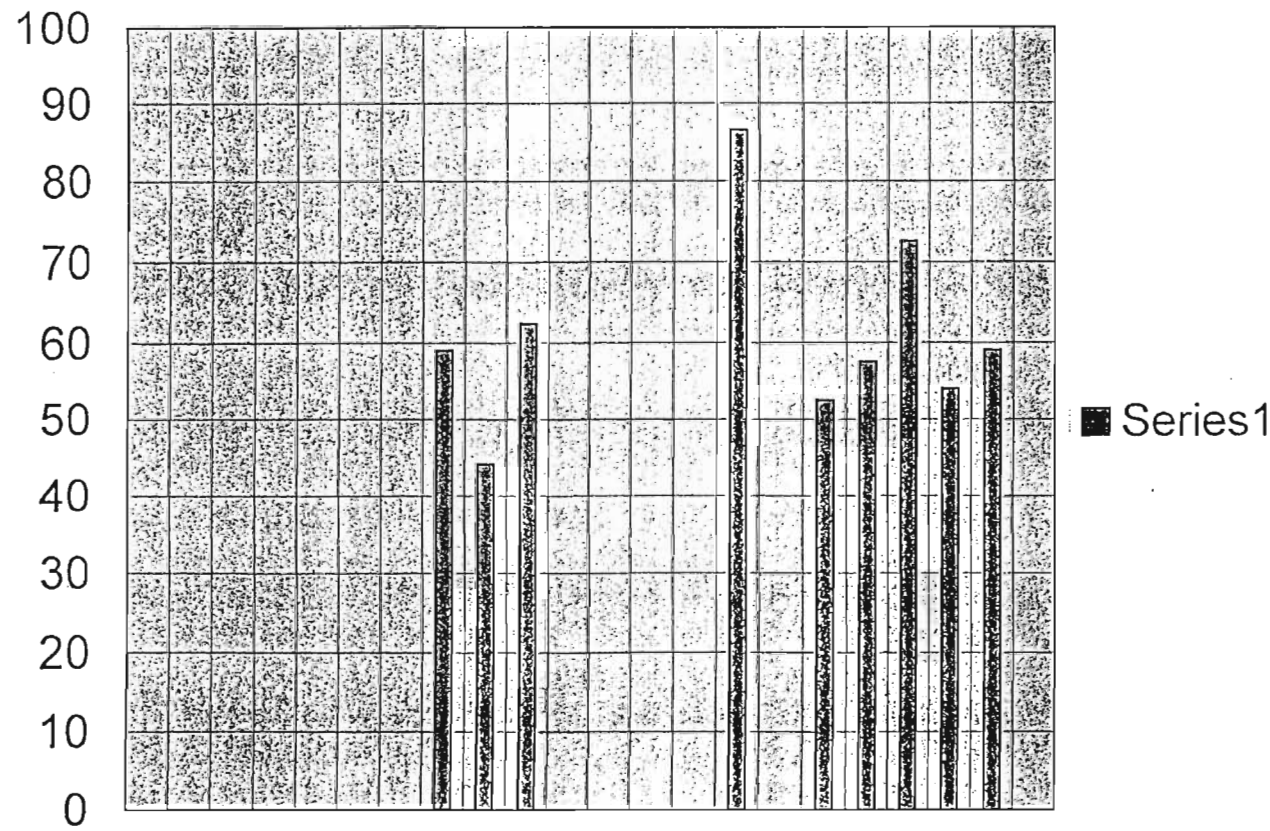
5 ppm Binned Spectrum of Pure Compound XVII



5 ppm Binned Spectrum of Impure Compound XVII



10 ppm Binned Spectrum of Pure Compound XVII



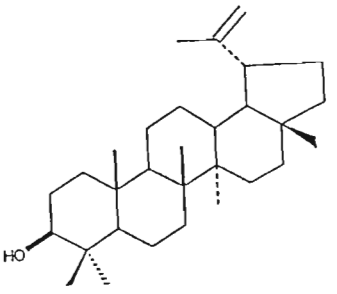
10 ppm Binned Spectrum of Impure Compound XVII

^{13}C NMR Data for Impure Compound XVIII

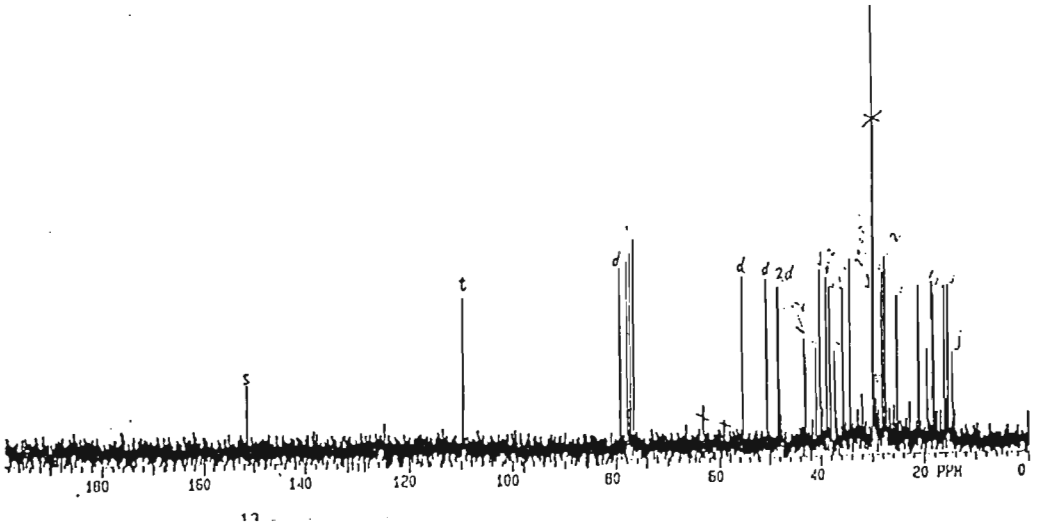
ppm	Intensity
151.403	25.479
109.649	58.02
79.229	72.074
• 77.887	72.074
• 77.25	72.074
• 76.612	72.074
55.43	66.635
50.552	66.517
48.41	60.99
48.111	62.351
43.113	40.791
42.931	27.68
40.92	36.83
40.104	67.547
38.956	38.583
38.795	67.212

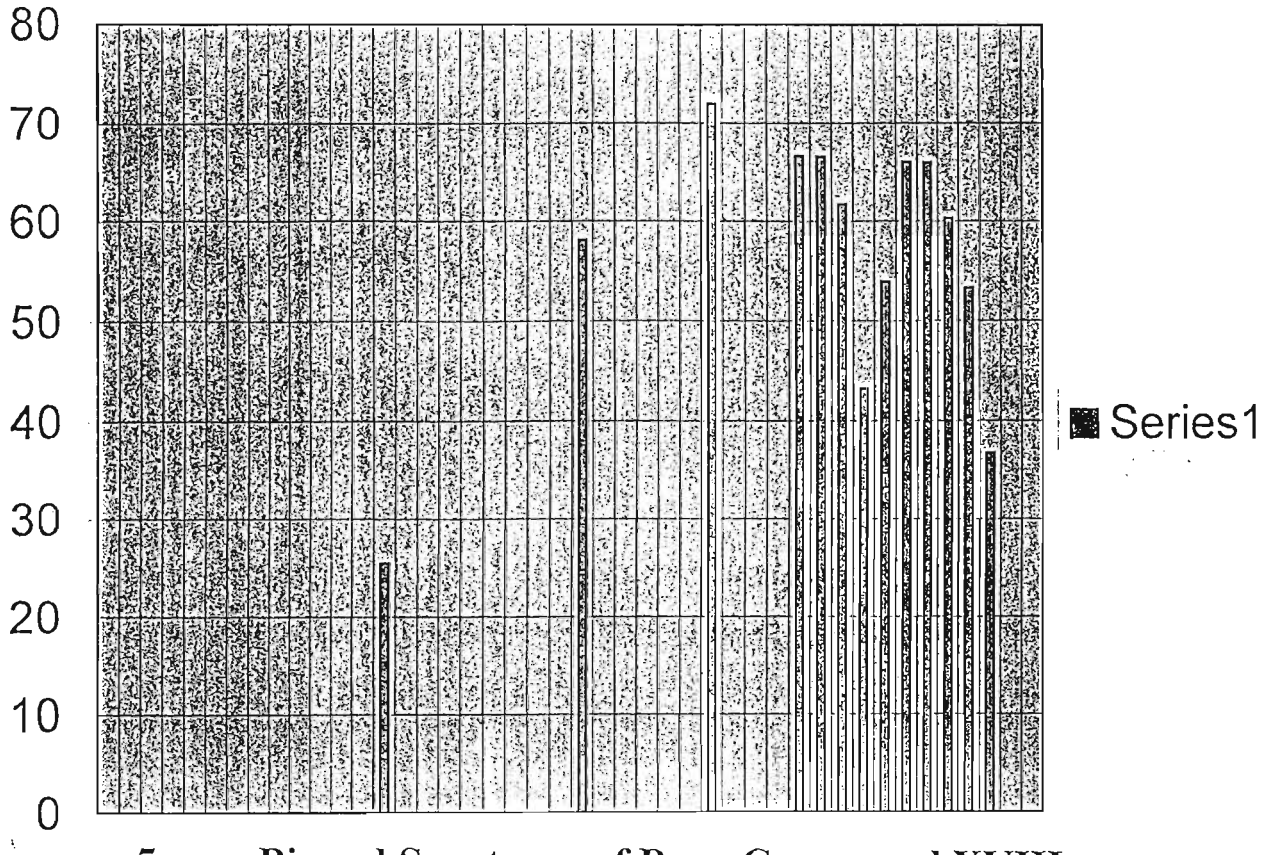
ppm	Intensity
38.13	63.218
37.249	37.669
35.667	62.68
34.351	65.968
29.908	64.39
• 29.789	175.637
• 29.704	40.385
28.061	67.335
27.502	68.766
27.455	71.353
25.176	57.841
20.969	60.264
19.352	37.075
18.357	61.637
18.047	56.43
16.162	60.676
16.007	42.631
15.42	61.214
14.577	36.761

• - denotes extraneous peaks

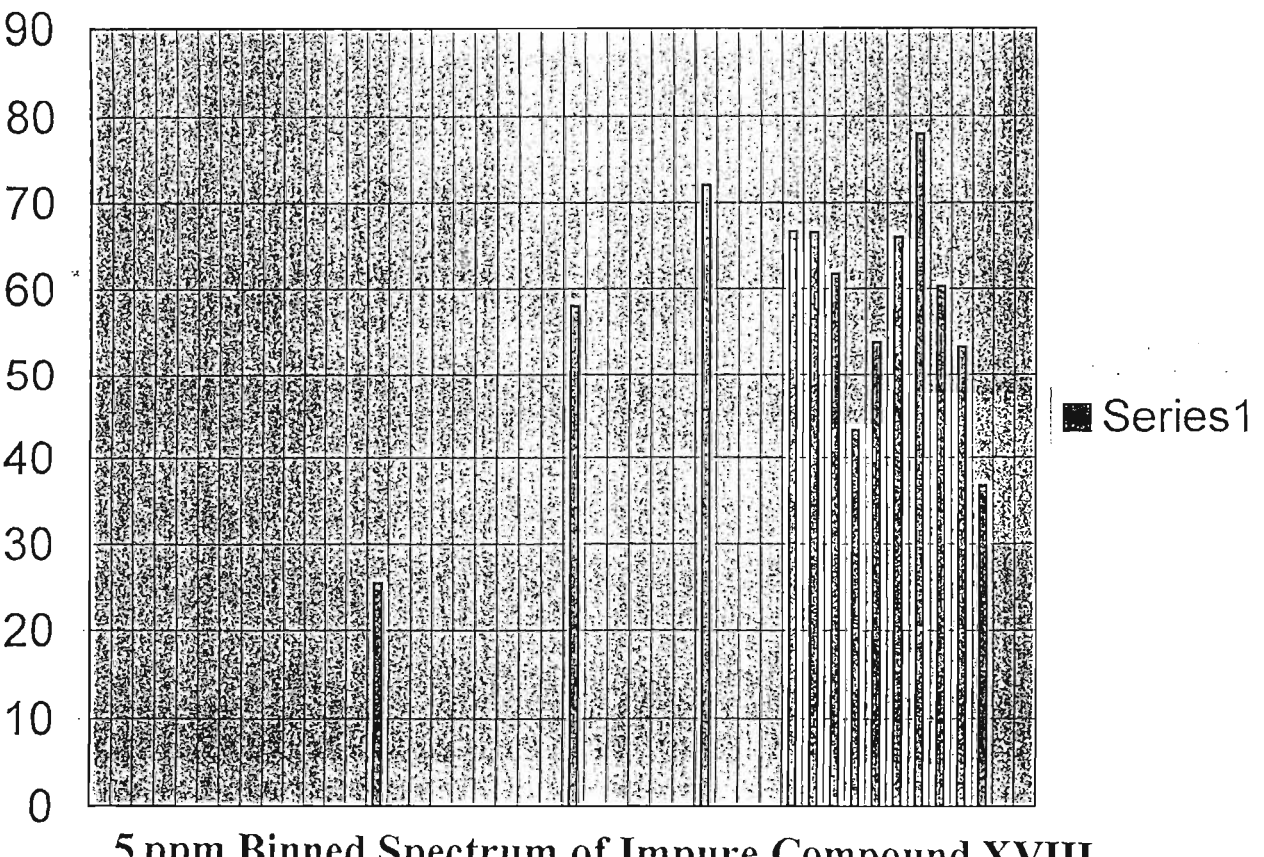


LUPEOL (XVIII)

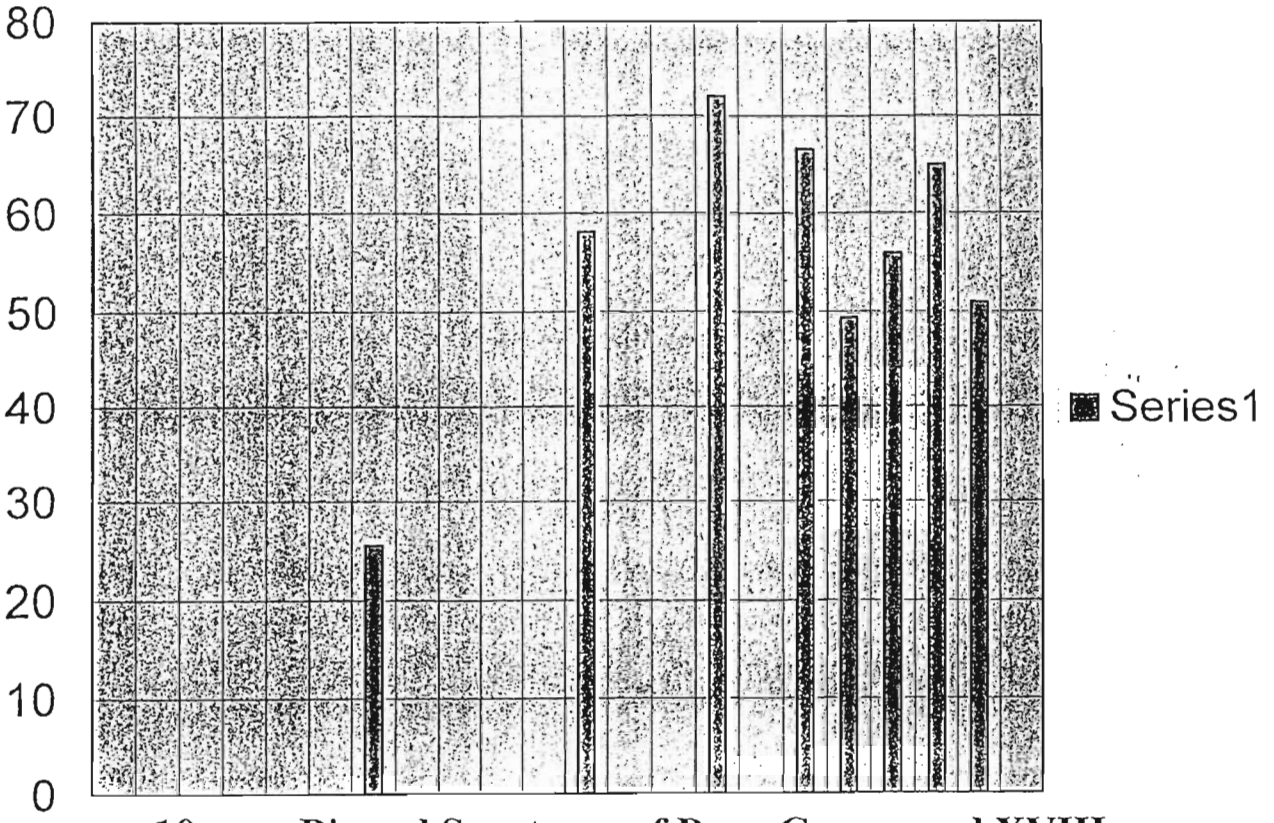




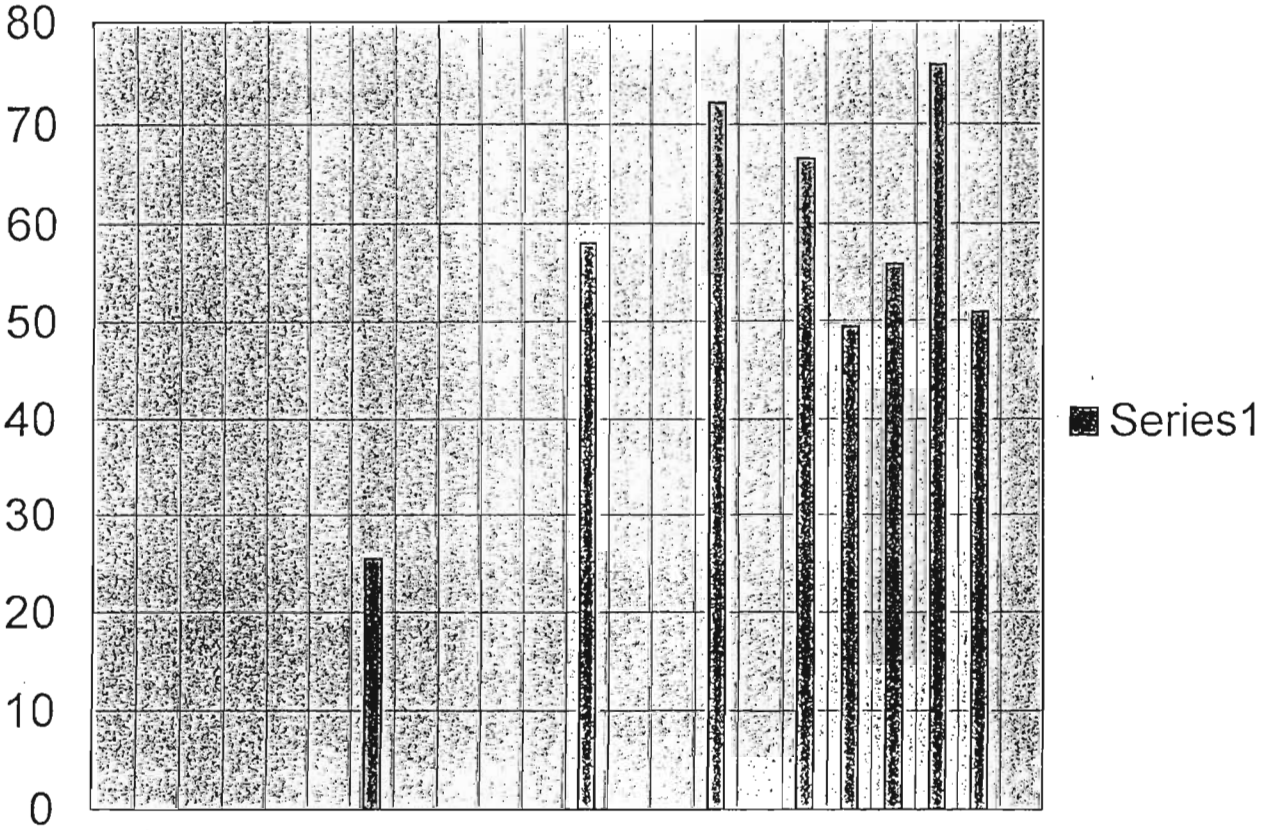
5 ppm Binned Spectrum of Pure Compound XVIII



5 ppm Binned Spectrum of Impure Compound XVIII



10 ppm Binned Spectrum of Pure Compound XVIII



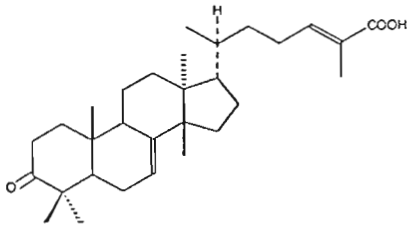
10 ppm Binned Spectrum of Impure Compound XVIII

^{13}C NMR Data for Impure Compound XXXXI

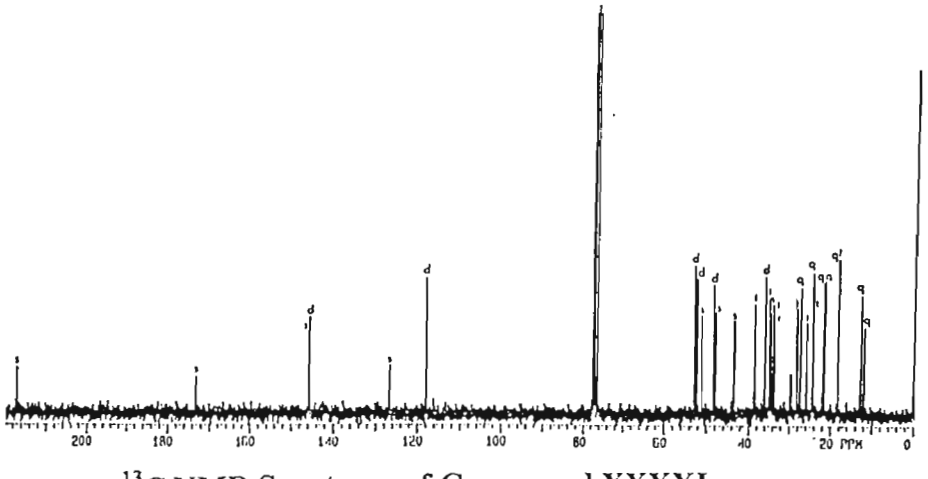
ppm	Intensity
217.064	19.874
173.141	15.261
145.83	36.344
145.659	41.891
126.667	20.125
117.876	59.246
• 77.211	10.676
52.886	64.744
52.288	58.227
51.161	43.017
48.428	55.941
47.884	43.795
43.526	39.982
38.522	47.386
• 36.103	10.362
36.02	59.451

ppm	Intensity
35.005	52.22
34.936	47.493
34.592	43.69
34.002	47.587
33.627	45.004
• 29.707	17.185
28.202	46.355
27.413	54.404
25.98	40.109
24.52	60.785
24.347	45.888
21.972	55.699
21.611	58.261
18.269	52.447
18.888	66.874
12.794	51.908
11.994	37.676
• 0	150.976

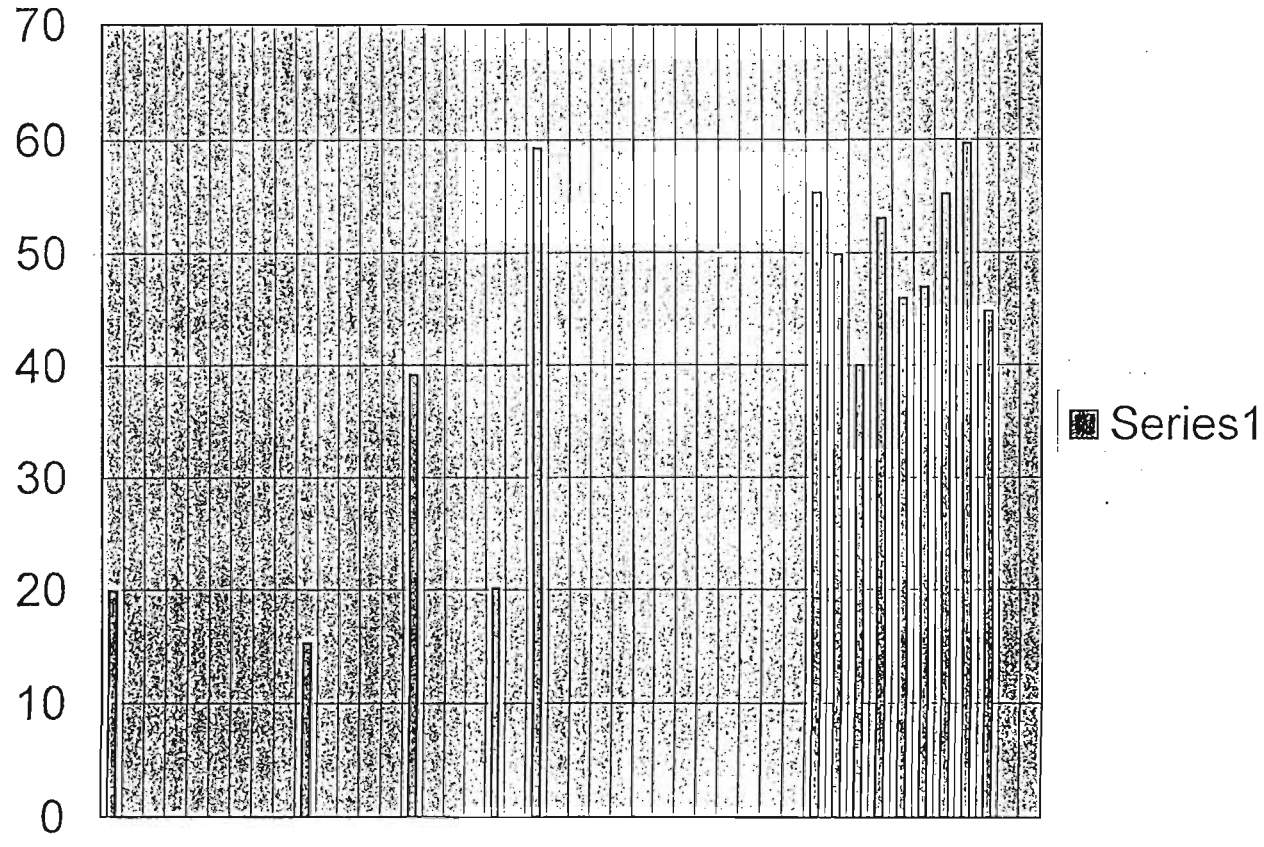
• - denotes extraneous peaks



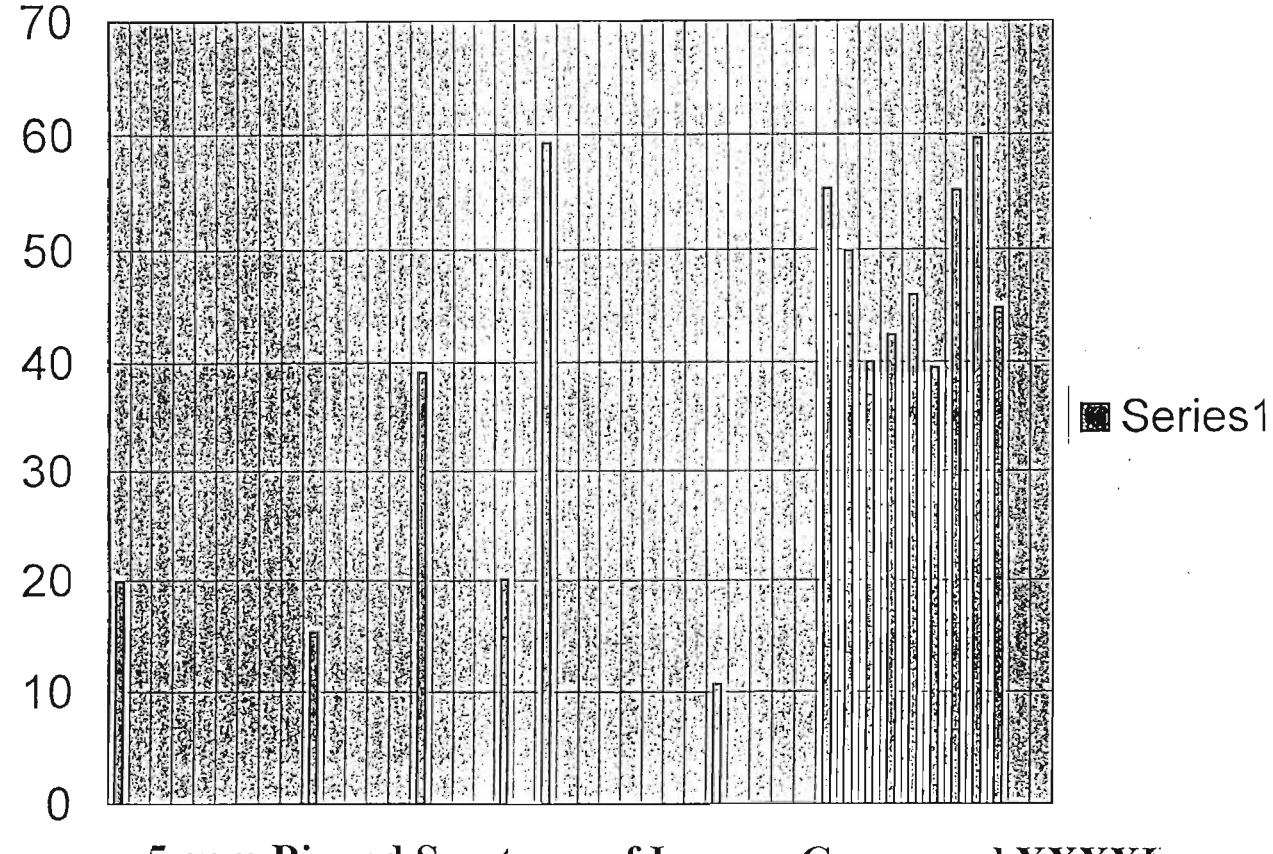
COMPOUND XXXXI



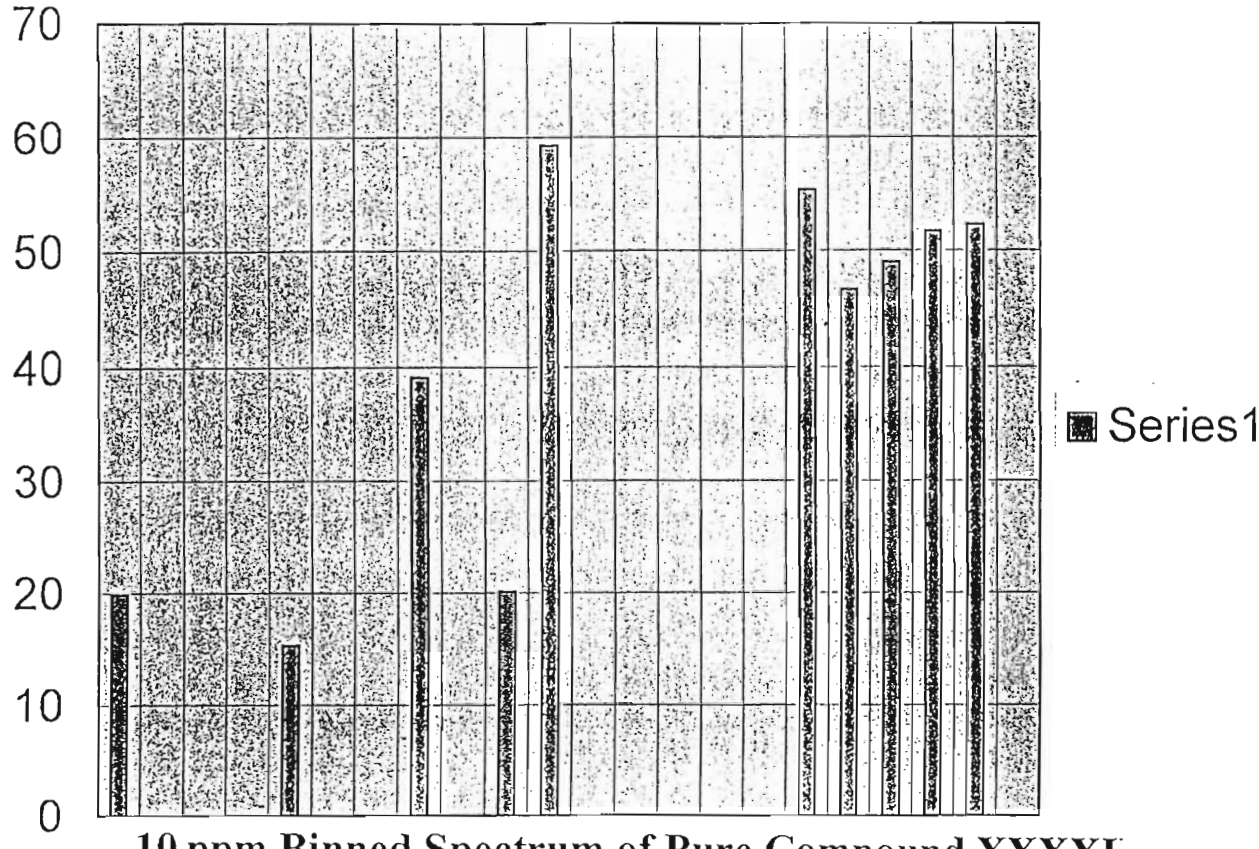
^{13}C NMR Spectrum of Compound XXXXI



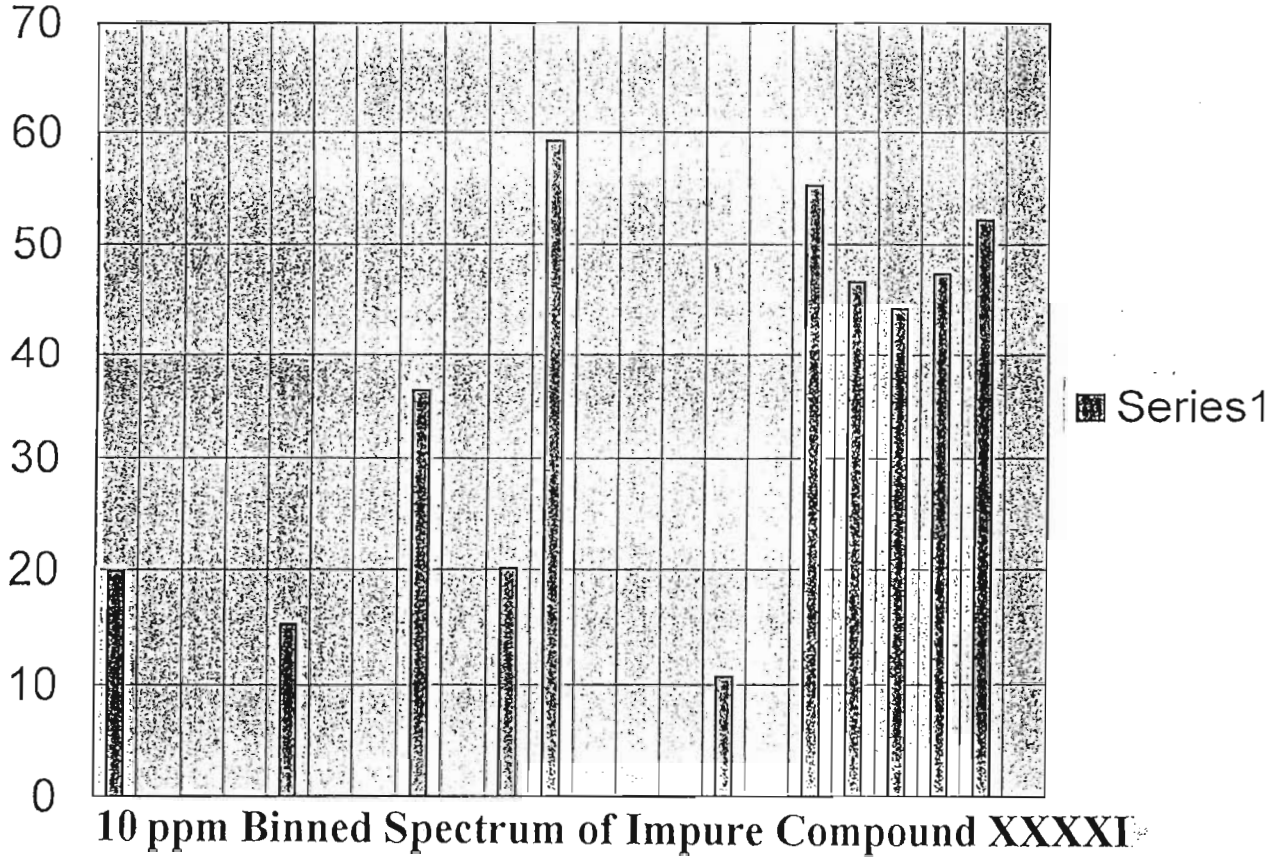
5 ppm Binned Spectrum of Pure Compound XXXXI



5 ppm Binned Spectrum of Impure Compound XXXXI



10 ppm Binned Spectrum of Pure Compound XXXXI



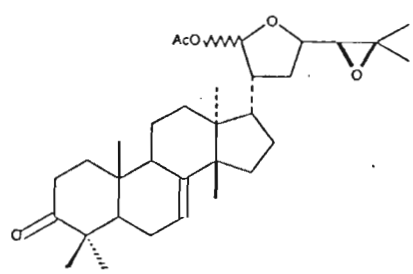
10 ppm Binned Spectrum of Impure Compound XXXXI

^{13}C NMR Data for Impure Compound X

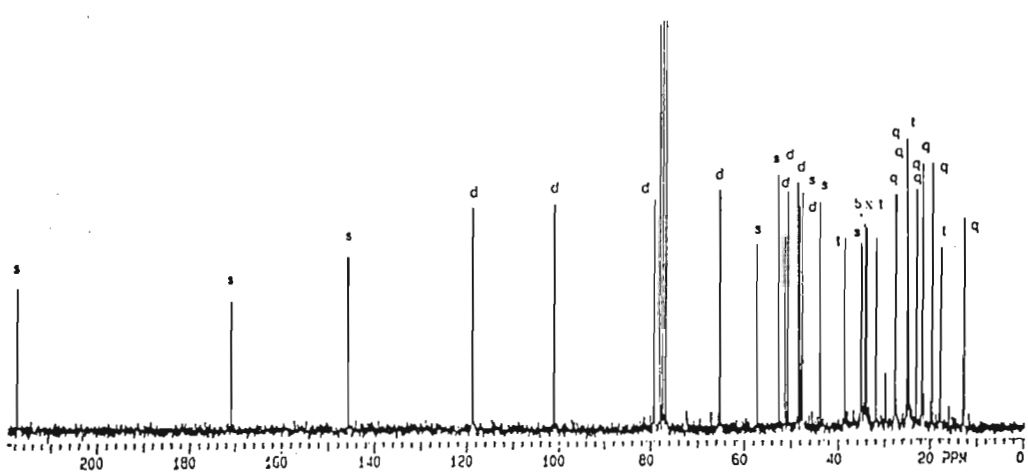
ppm	Intensity
217.46	53.836
170.879	50.701
145.778	69.449
118.684	87.104
100.978	88.546
79.167	88.013
• 77.937	88.013
• 77.3	88.013
• 76.663	88.013
65.132	94.319
57.207	76.804
52.527	97.603
51.15	72.63
50.494	94.967
48.413	93.021
48	88.561
47.496	89.775

ppm	Intensity
43.736	85.57
38.568	71.98
35.197	85.515
35	68.815
34.384	77.269
33.912	76.328
31.89	71.821
• 29.798	20.165
27.514	76.545
27.402	91.852
24.994	109.847
24.606	95.684
24.45	72.051
22.902	90.804
21.647	100.254
21.485	71.214
19.58	100.397
17.796	68.295
12.78	82.276

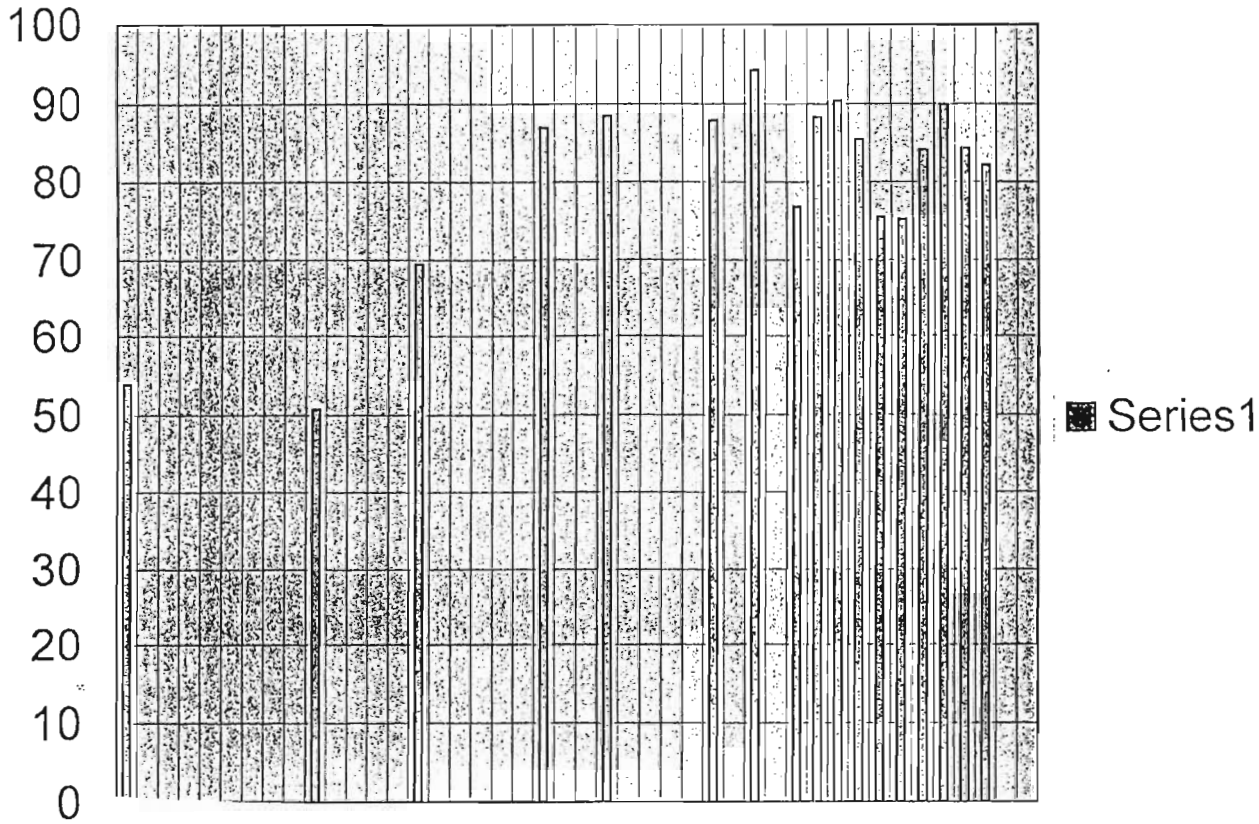
• - denotes extraneous peaks



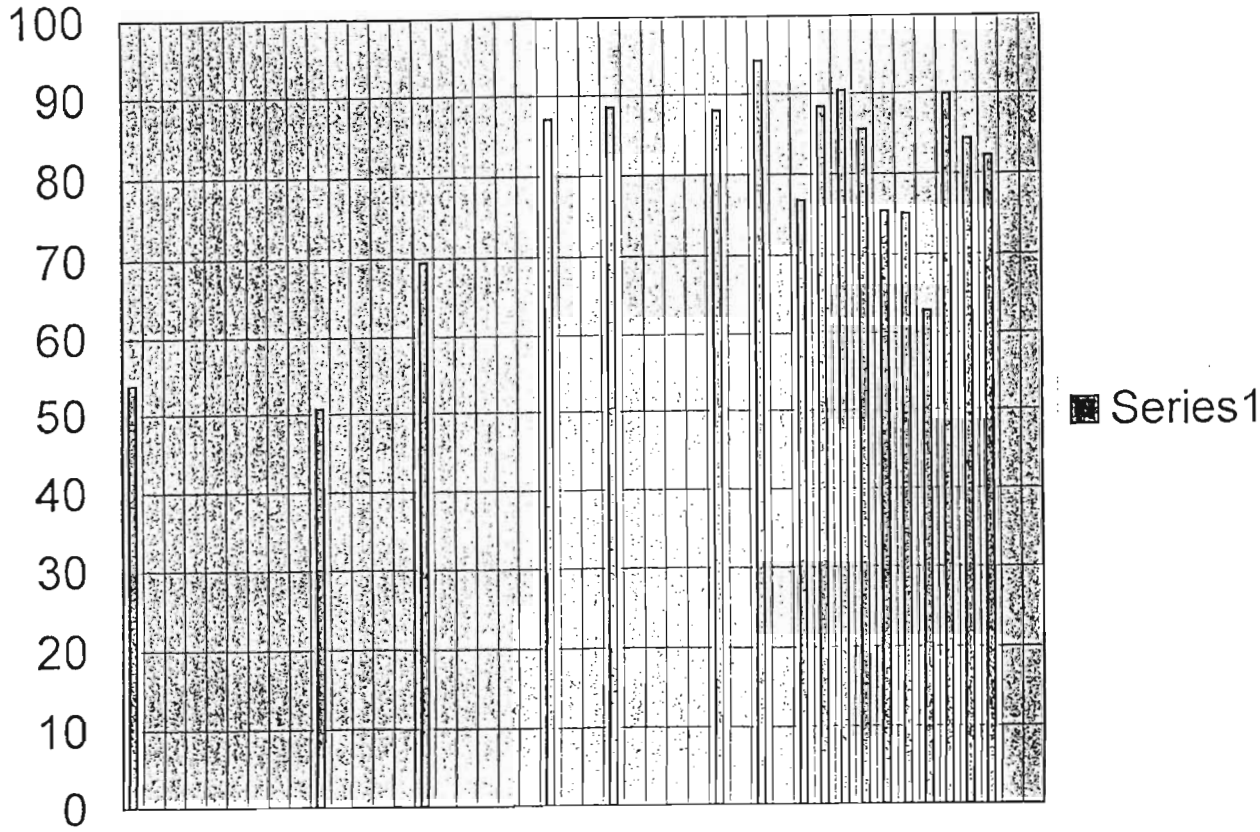
MELIANONE ACETATE (X)



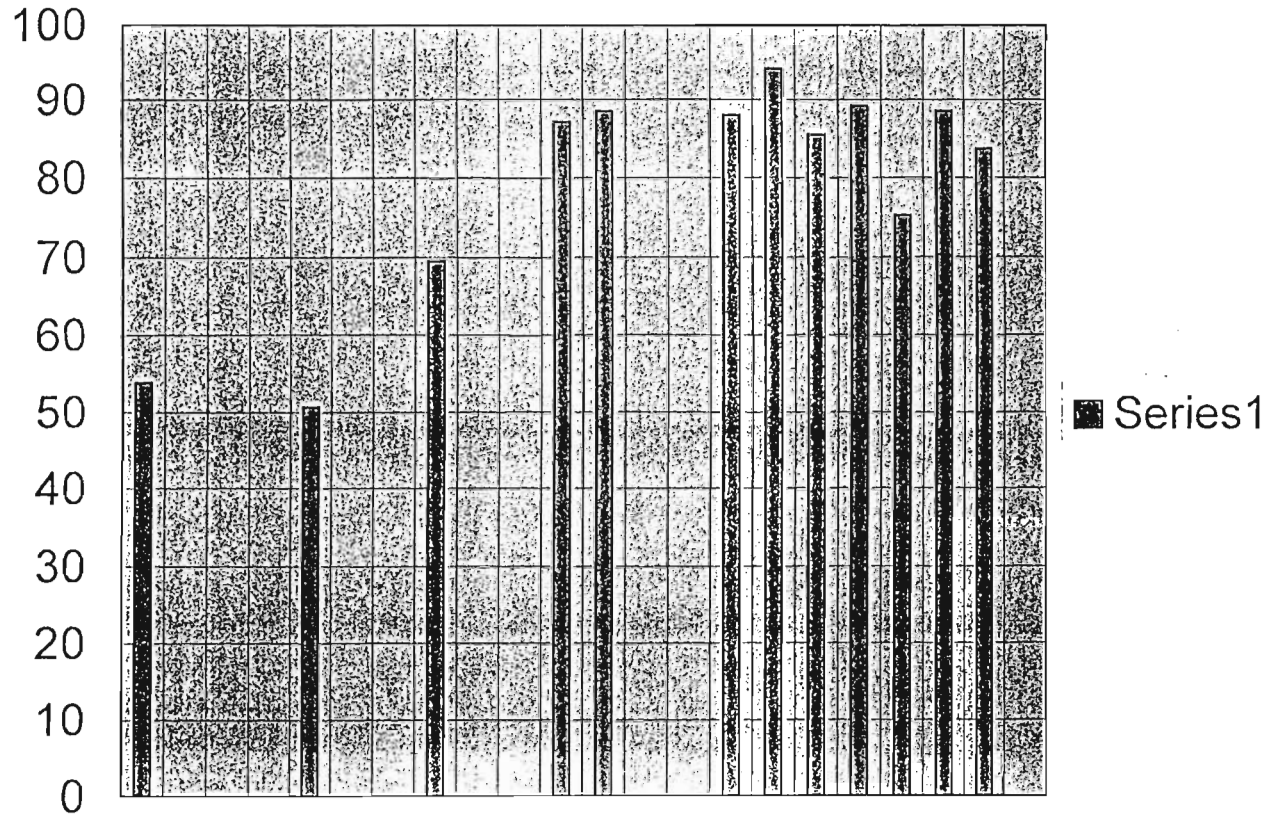
^{13}C NMR Spectrum of Compound X



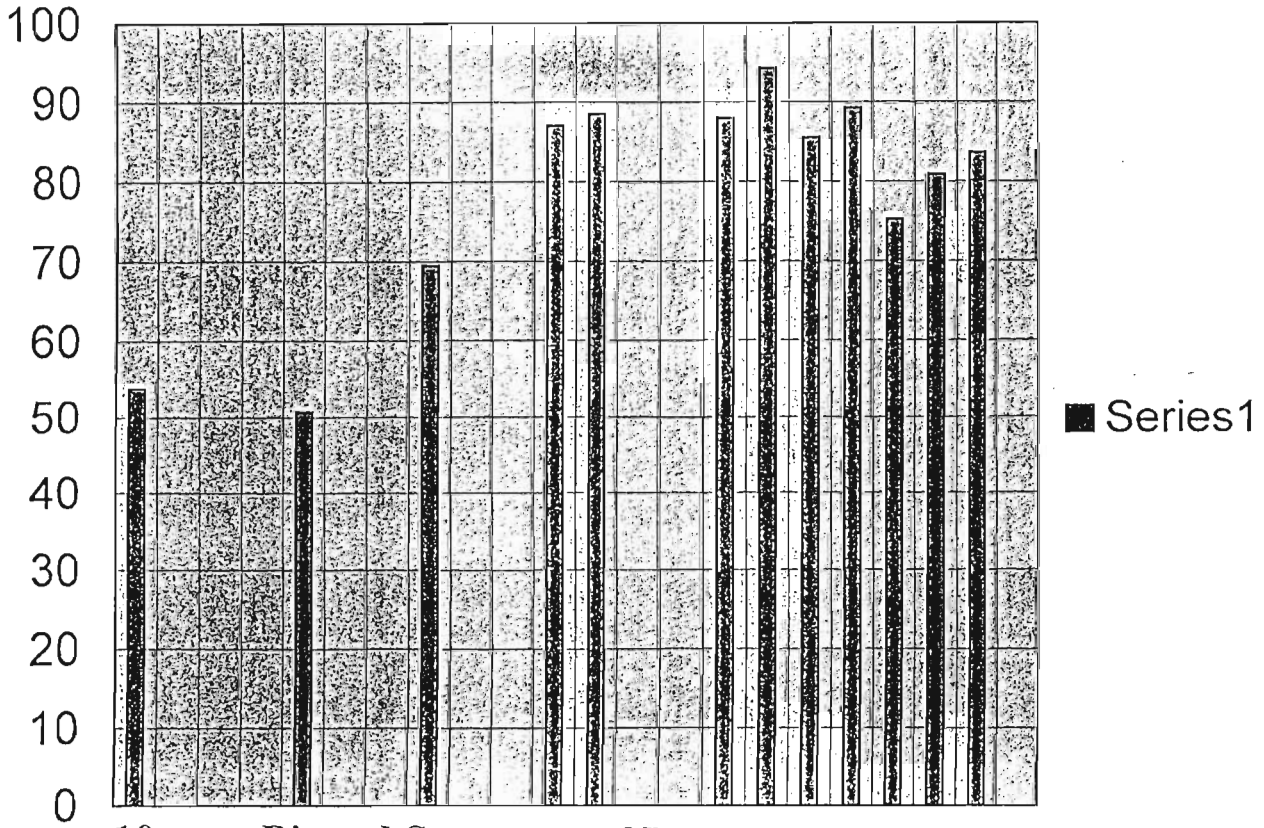
5 ppm Binned Spectrum of Pure Compound X



5 ppm Binned Spectrum of Impure Compound X



10 ppm Binned Spectrum of Pure Compound X



10 ppm Binned Spectrum of Impure Compound X