

A new automated assign and analysing method for high-resolution rotationally resolved spectra using genetic algorithms

W Leo Meerts¹ and Michael Schmitt²

¹ Molecular- and Biophysics Group, Institute for Molecules and Materials, Radboud University Nijmegen, PO Box 9010, Nijmegen 6500 GL, The Netherlands

² Institut für Physikalische Chemie, Heinrich-Heine Universität Düsseldorf, Gebäude 26.43.02, Universitätsstraße 1, Düsseldorf 40225, Germany

E-mail: leo.meerts@science.ru.nl

Received 14 May 2005

Accepted for publication 15 August 2005

Published 21 December 2005

Online at stacks.iop.org/PhysScr/73/C47

Abstract

This paper describes a numerical technique that has recently been developed to automatically assign and fit high-resolution spectra. The method makes use of genetic algorithms (GA).

The current algorithm is compared with previously used analysing methods. The general features of the GA and its applications in automated assignments is discussed. In a number of examples the successful application of the technique is demonstrated.

PACS numbers: 07.05.Kf, 33.20.Lg, 33.20.Sn, 33.15.Mt, 36.40.Mr

1. Introduction

Until a couple of years ago spectroscopists all over the world were mainly using traditional manners of spectral assignments. The methods are based on finding regularities and performing a *by eye* pattern recognition and assign in this way quantum numbers to the transitions. In general this is a tedious process and even a relative simple spectrum as for example of the rotationally cooled naphthalene molecule [1] can take an experienced scientist from a couple of days up to several weeks. If the analysis involves a series of spectra for example as function of vibrational quantum number or from different isotopomers or conformers the amount of work rapidly grows out of hand. Consequently, the traditional techniques inhibit the study of larger and more complicated systems such as molecules of biological interest.

A number of tools has been developed that make use of fast computers and their graphical possibilities to facilitate the assignment *by eye*. Very widely used and with great success, particularly in the analysis of microwave spectra, is the Loomis–Wood method [2]. A Loomis–Wood diagram is a two-dimensional peak diagram in which the occurrence of a transition is plotted versus frequency, in segments, with successive segments displayed one above another. Such diagrams were first used by Loomis and Wood in 1928. Because of the time required to manually create a

Loomis–Wood diagram, they were not useful in the initial assignment of spectra before the advent of microcomputers. The first computer program to generate a Loomis–Wood plot was written at the Ohio State University in the 1960s [3] and the first interactive Loomis–Wood applications by Winnewisser *et al* [4] appeared in the 1980s. Loomis–Wood programs are particularly useful for the analysis of congested spectra of symmetric tops, slightly asymmetric tops, and linear molecules. Neese [5] has developed an interactive Loomis–Wood assignment package. A nice example of a recent application of the Loomis–Wood method can be found in the paper of Thompsom *et al* [6].

In this frame the computer program *JB95* developed by Plusquellic [7] should be mentioned. This program has been very successful in the graphical assignment (*by eye*) of high-resolution rotationally resolved spectra. The program consists of a graphical user interface based on a Windows platform.

In a recent paper, Morruzi [8] presented an investigation on the feasibility of automated molecular line assignment. Dense rovibrational molecular spectra are normally assigned by strongly interactive computer methods, ranging from commercial spreadsheets to dedicated programs, like Loomis–Wood or Ritz. While a general-purpose, fully automated assignment procedure seems to be out of reach for the near future, he shows that a thorough investigation of the problem can lead to new, more efficient and less interactive methods,

at least in reasonably favourable conditions. Interesting suggestions are provided by some modern *heuristic* problem-solving algorithms, which mimic natural processes.

In order to try to solve the assignment problems with the help of a computer the group of Neusser [9] has developed a procedure, which directly fits the experimental data, without prior assignments. This method, which is called ‘correlation automated rotational fitting’, has been pioneered by Levy and co-workers [10–12], and uses the correlation between the experimental and the simulated spectrum as a measure of the quality of the fit. Unfortunately, the method still has limited applicability.

The outcome of a first study in which genetic algorithms (GA) were used to solve the automatic assignment problem was very promising and resulted in a paper by Hageman *et al* [13]. In that paper it was shown that for a series of previously manually assigned spectra of molecules like indole, indazole, benzimidazole [14] and 4-aminobenzonitrile [15], an automatic fitting based on GA was successful. A crucial role in the success was the development of a proper fitness function.

In a further series of papers Meerts and Schmitt showed that the GA method to automate the assignment of complicated and entangled spectra was extremely successful [16–22].

Related methods using GA have previously been used in a variety of other spectroscopic applications such as Nuclear Magnetic Resonance [23], fluorescence/absorption spectra in polyatomic molecules [24], Mössbauer spectroscopy [25], x-ray spectra from plasmas [26] and powder EPR spectra [27].

In this work we discuss the general GA method, its application and some of the highlights and successes. The examples given in this paper are from high-resolution gas phase spectra. However, the discussed automated assignment with the help of the GA method can be applied on a much wider range of spectra.

2. The genetic algorithm

A description of the GA used in this investigation can be found in [13, 18]. The GA is basically a global optimizer, which uses concepts copied from natural reproduction and selection processes. For a detailed description of the GA the reader is referred to the original literature [28–30]. We shortly introduce the elements of the GA, which will be used in the following.

- Representation of the parameters: the molecular parameters are encoded binary or as real data type, each parameter representing a gene. A vector of all genes, which contains all molecular parameters is called a chromosome. In an initial step the values for all parameters are set to random values between lower and upper limits which have to be chosen by the user. No prior knowledge of the parameters is necessary. A total of 300–500 chromosomes are randomly generated, forming a population.
- The solutions (chromosomes) are evaluated by a fitness function (or cost function), which is a measure for the quality of the individual solution. The fitness function which is used here, is described below.

- One optimization cycle, including evaluation of the cost of all chromosomes is called a generation. Generally, convergence of the fit in our case is reached after 300–500 generations.
- Pairs of chromosomes are selected for reproduction and their information is combined via a crossover process. This crossover might take place as a one-point, two-point or uniform crossover. A crossover just combines information from the parent generations. It basically explores the error landscape.
- The value of a small number of bits is changed randomly. This process is called mutation. Mutation can be viewed as exploration of the cost surface. The best solutions within a generation are excluded from mutation. This elitism prevents already good solutions from being degraded.

The performance of the GA depends on internal parameters like mutation rate, elitism, crossover probability and population size, which therefore should also be optimized for a given problem. Fortunately this meta-optimization results in similar parameters for quite different problems of optimization. The meta-optimization for some of the parameters is described in [18].

A proper choice of the fitness function is of vital importance for the success of the GA convergence. In [13] and [18] the fitness function F_{fg} has been defined as

$$F_{fg} = \cos(\alpha) = \frac{(\mathbf{f}, \mathbf{g})}{\|\mathbf{f}\| \|\mathbf{g}\|}. \quad (1)$$

In this equation \mathbf{f} and \mathbf{g} are the vector representation of the experimental and calculated spectra, respectively. The inner product (\mathbf{f}, \mathbf{g}) , defined with the metric \mathbf{W} with matrix elements $W_{ij} = w(|j - i|)$, has the form

$$(\mathbf{f}, \mathbf{g}) = \mathbf{f}^T \mathbf{W} \mathbf{g}, \quad (2)$$

and the norm of \mathbf{f} is $\|\mathbf{f}\| = \sqrt{(\mathbf{f}, \mathbf{f})}$; similar for \mathbf{g} . For $w(|j - i|)$ we used a triangle function [13] with a width of the base of Δw

$$w(j - i) = \begin{cases} 1 - |j - i| / (\frac{1}{2} \Delta w) & \text{for } |j - i| \leq \frac{1}{2} \Delta w \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

Important but not decisive is the reliability of the experimental intensities and the presence of a model capable to explain the observed spectra. These conditions can however be released significantly in some cases as will be demonstrated below.

3. Examples of the success of GA automatic assignments

The experimental rotationally resolved spectra are obtained from a high-resolution UV laser spectrometer. Its resolution is of the order of 0.001 cm^{-1} or 25 MHz (full width at half maximum, Δ_{FWHM}) at $35\,000 \text{ cm}^{-1}$. This is achieved by crossing a single-frequency continuous wave laser with a molecular beam and detecting the laser induced fluorescence. The details are described elsewhere [1, 31, 32].

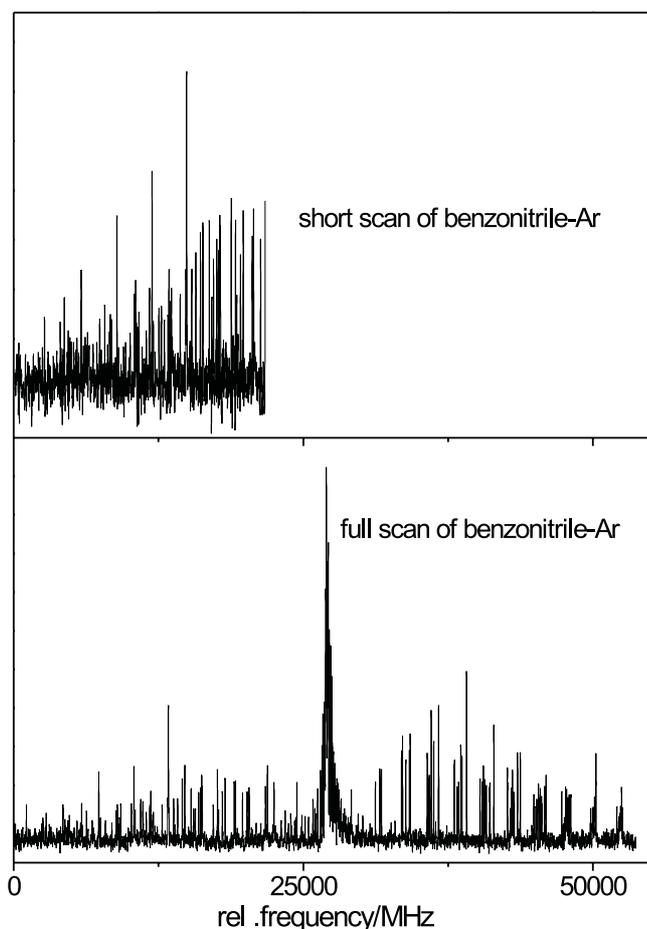


Figure 1. Upper trace: low frequency part of the spectrum of benzonitrile-Ar. Lower trace: complete rovibronic spectrum. Intensities are given in arbitrary units. For details see text.

For the simulation of the rovibronic spectra a rigid asymmetric rotor Hamiltonian was employed [33]. The parameters to be determined by the GA are the three rotational constants A , B and C for each electronic state, the origin frequency ν_0 of the vibronic band and line intensity determining parameters like rotational temperature, transition dipole moment orientation and line shape parameters. The relative intensities were fit to a two-temperature model [18, 34]. The GA software used was the library package PGAPack version 1.0 [35]. This package performs excellently on parallel processor systems. Most calculations were performed on a dual processor PC with two Pentium 2.8 GHz processors under Linux. Typical computing times on this system are 8 min wall clock time for a full GA fit of a single spectrum³.

3.1. GA fit of very dense rovibronic spectra

In the following we will present an automated GA fit of a rovibronic spectra, which is very dense due to small rotational constants. These kind of spectra normally do not represent a great difficulty for the GA, as will be shown below.

³ On request, the authors make available a full version of their GA-program. The program has been thoroughly tested under Linux as well several UNIX versions. The package contains the program and an extensive manual with the installation procedure. License conditions are applicable if the program is used.

Table 1. Molecular constants from a GA assignment of the partial spectrum of the origin of benzonitrile-Ar, the complete spectrum and an assigned fit. See text for details. The orientation of the dipole moment vector is determined by the polar angle θ and the azimuthal angle ϕ . The rotational temperature T has been determined from the relative intensities of the transitions in the spectrum. The origin of the spectrum is at $36489.04(2) \text{ cm}^{-1}$.

Parameter	GA fit ^a	GA fit ^b	Assigned
A''/MHz	1343.80(150)	1347.32(19)	1347.58(18)
B''/MHz	1002.55(121)	1004.99(4)	1004.98(14)
C''/MHz	717.68(108)	718.99(4)	717.70(39)
$\theta/^\circ$	22(3)	17.53(7)	20
$\phi/^\circ$	82(5)	70.05(2)	70
T/K	1.71(3)	1.68(3)	2
$\Delta A/\text{MHz}$	-32.89(44)	-32.61(3)	-32.47(23)
$\Delta B/\text{MHz}$	20.60(30)	21.08(16)	20.87(14)
$\Delta C/\text{MHz}$	6.25(15)	6.76(7)	6.80(34)

^aFit to the spectrum in the upper trace of figure 1.

^bFit to the spectrum in the lower trace of figure 1.

As an example, and a good demonstration of the power of the automated fitting procedure, we discuss here the benzonitrile-Ar spectrum. If due to experimental limitations only the outermost parts of the P - or the R -branch can be recorded and the electronic origin of a rovibronic band is missing, the task of performing an assigned fit gets tedious or even impossible. However, also in this difficult case the GA succeeds in finding the global minimum and assigning the spectrum properly. We chose the spectrum of the electronic origin of benzonitrile-Ar, shown in the upper trace of figure 1 to demonstrate this. Obviously only the low-frequency side of the spectrum has been measured with a quite bad signal-to-noise ratio. Nevertheless, the GA was able to determine the molecular parameters. The result is given in the first column of table 1. A GA fit to the complete spectrum with good signal-to-noise (lower trace in figure 1) yields slightly different molecular parameters (second column of table 1). Nevertheless, the quality of the parameters obtained from the fit to the partial spectrum is surprisingly good.

3.2. Simultaneous GA fits of a number of overlapping rovibronic spectra

A much more demanding task than a fit of a single rovibronic spectrum is the simultaneous fit of two (or more) overlapping spectra. Firstly, the number of transitions within a spectral interval is multiplied, leading to very dense and congested spectra. Secondly, the number of molecular parameters is also multiplied, which generates quite a large parameter space. Below we show that the GA-spectrum assignments are capable to handle overlapping spectra from different isotopomers.

We performed a fit [18] of the rovibronic spectrum of the isotopomeric pair benzonitrile-²⁰Ne/benzonitrile-²²Ne in the natural abundance of ²⁰Ne/²²Ne (91 : 9), shown in figure 2. In this case the GA has the very difficult task of fitting quite a weak spectrum in the presence of a strong spectrum. The situation is further complicated by the fact that some of the lines present in the spectrum are due to benzonitrile monomer

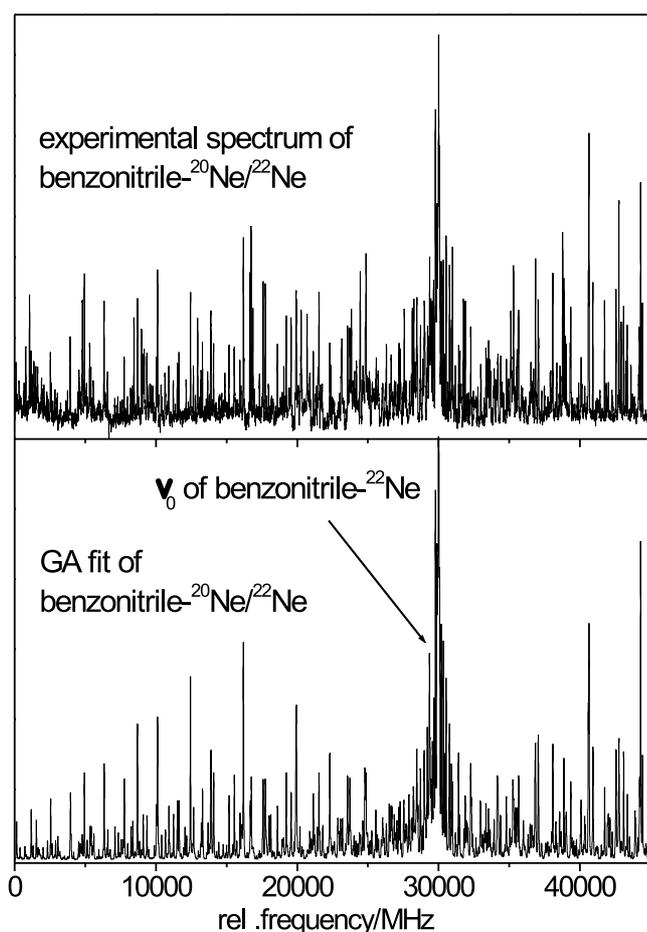


Figure 2. Upper trace: experimental spectrum of benzonitrile-²⁰Ne/benzonitrile-²²Ne. Lower trace: simulation using the best parameters from [18]. For signal-to-noise reasons it looks as if the intensity patterns do not fully match. A second reason is the presence of high J -state transitions of the monomer in this region.

lines (the electronic origin of the monomer is shifted by about 4.3 cm^{-1} to higher frequency). Although the monomer origin has already been assigned [36], these monomer lines cannot be predicted with sufficient accuracy, because they belong to very high J -states.

The molecular parameters were obtained from a four step GA fit. In the first step $\Delta w/\Delta_{FWHM} = 10$ was employed. The search limits for the rotational constants were $\pm 100\text{ MHz}$ for both isotopomers. The parameter limits were narrowed down to one-tenth of the original size, centred around the best fit value of the first step. While the more abundant species (benzonitrile-²⁰Ne) presented no difficulties, the fit of the weaker component spectrum got trapped in a local minimum. This had two reasons: the intensity of the sub-spectrum of benzonitrile-²²Ne is only one-tenth of the stronger component and the additional monomer lines have comparable intensities to the transitions of the stronger isotopic species. Thus, the parameter limits for the weaker sub-spectrum had to be reduced more slowly and in more steps. Firstly, only by a factor of two, while $\Delta w/\Delta_{FWHM}$ was reduced to 7.5. In a subsequent step $\Delta w/\Delta_{FWHM} = 5$ and limits of $\pm 20\text{ MHz}$ for the rotational constants were employed. Finally, the molecular constants were obtained for $\Delta w/\Delta_{FWHM} = 1.5$.

In this case the fit required quite some ‘fine tuning’ which had to be done manually. Nevertheless, the results of the fit

of the rovibronic spectrum of benzonitrile-²⁰Ne/benzonitrile-²²Ne show that even very congested spectra, with one spectral component much weaker than the other can be assigned using the GA without any prior knowledge of geometry or molecular parameters.

The spectrum of 7-azaindole (7AI) was automatically [19] assigned using the GA-based fit. The rotational constants obtained from this GA fit are reported in table 2.

Soft deuteration, by adding 20 mbar D₂O to the Ar seed gas prior to expansion resulted in the spectrum shown in figure 3 of [19]. Apparently a second band emerges, which can be assigned to the 7AI[ND] isotopomer. Both bands were fit together using the GA. The molecular constants of the first spectrum, like rotational constants, origin frequency, and Lorentzian width were set fixed, while the global parameters for the complete fit like temperature(s), weights, baseline, relative intensity of both spectra etc. were allowed to vary. The resulting rotational constants for 7AI[ND] for both electronic states are listed in table 2.

Higher deuteration grades were obtained by three times refluxing 7AI with an excess of DCl in D₂O (38%) and a subsequent removal of the solvent. This resulted in a 50:50 mixture of mono- and bi-deuterated species. Figure 3 shows the resulting high-resolution spectrum. Two new bands appeared to the blue of the two isotopomers already described. From a mass spectrum of the deuterated substance we know, that the highest deuteration grade was d₂. Thus, one of the new bands belongs to a d₂-isotopomer, the other one to a d₁-isotopomer, distinct from 7AI[ND]. We fitted the complete spectrum with the parameters of the first two bands kept fixed to the above determined parameters. The GA succeeded in finding the rotational constants for the other two isotopomers. Their values are given in table 2.

Since the GA performs a lineshape fit of the complete spectrum, much better information on the linewidth is gathered than from a lineshape fit to a few individual lines. In order to obtain the relevant parameters that determine the intensities in the spectrum, we performed a second GA fit with a reduced search range and the weight function width $\Delta w = 0$. This resulted in improved values for the angles θ and ϕ that are connected to the components of the transition dipole moment.

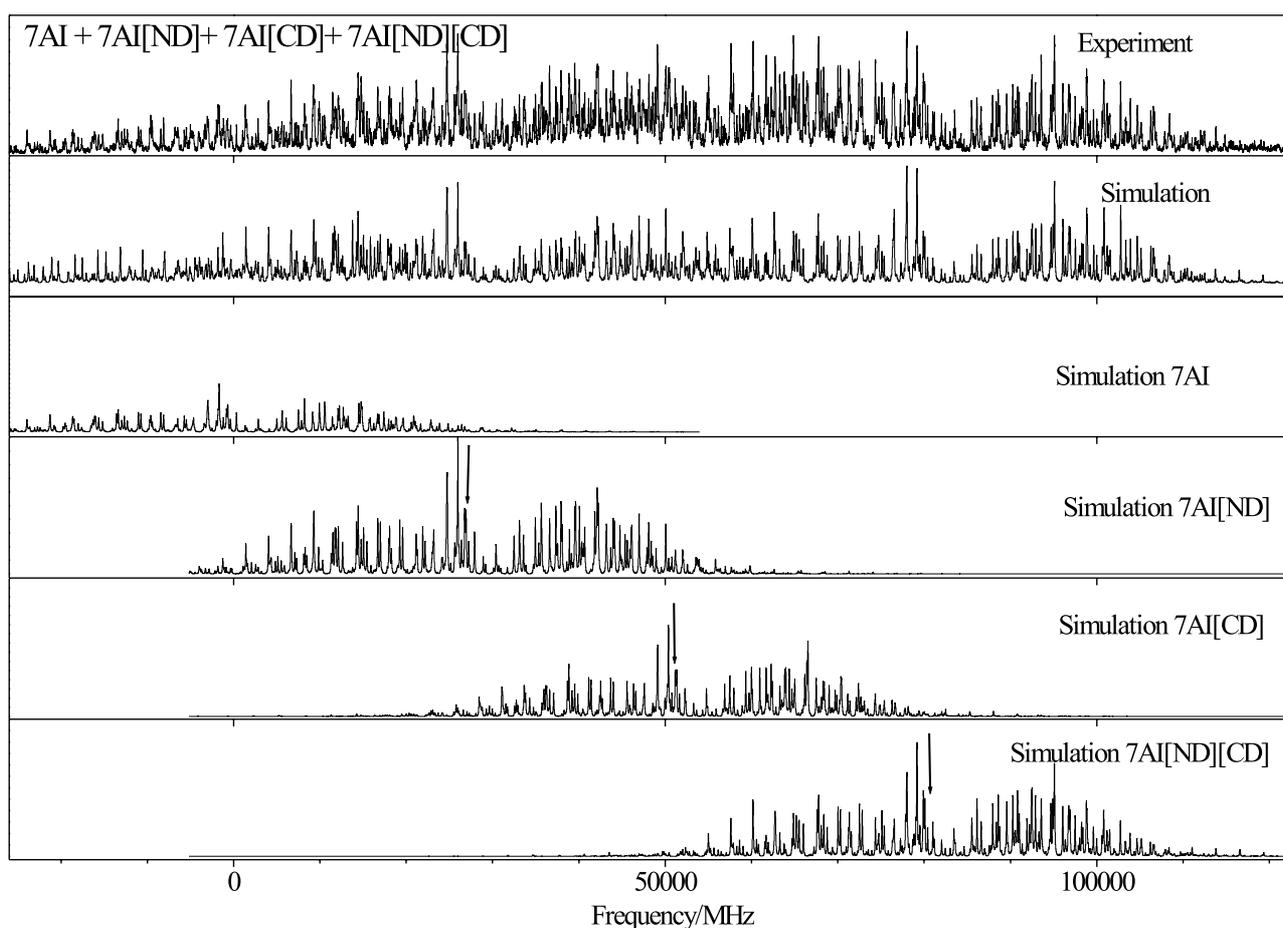
The determination of the Lorentzian component of the line width can be improved using the fit of all available intensities. With a fixed Gaussian contribution of 25 MHz from the experiment we obtained a Lorentzian contribution of $64 \pm 1\text{ MHz}$ for 7AI from a GA analysis. The resulting S_1 lifetime was $2.55 \pm 0.03\text{ ns}$. The lifetime of 7AI[ND] has been determined from the spectrum of two isotopomers given in figure 3 from [19] to be 2.34(2) ns, slightly shorter than of 7AI.

4. Summary

In this paper, we have shown that the GA is capable to treat a wide range of different spectra with complexity ranging from highly overlapping transitions to coinciding spectra of different isotopomers. Spectra that are life-time broadened can also be successfully analysed as long as they contain sufficient structure. Even if the signal-to-noise ratio is low

Table 2. Molecular parameters of the electronic origin band of 7-Azaindole as obtained from the GA fit. All values are given in MHz.

	7-AI	7-AI[ND]	7-AI[CD]	7-AI[ND][CD]
A''	3928.93(2) ^a	3807.60(3) ^a	3794.95(60)	3674.46(24)
B''	1702.629(3) ^a	1684.722(2) ^a	1678.73(16)	1662.45(15)
C''	1188.128(5) ^a	1168.241(2) ^a	1164.10(8)	1144.89(5)
ν_0^b	0	27553.67(12)	51934.78(98)	80640.38(39)
ΔA	-183.47(11)	-173.77(6)	-172.11(10)	-162.50(8)
ΔB	1.24(5)	-0.48(3)	-0.62(4)	-0.07(5)
ΔC	-16.62(3)	-16.95(2)	-16.89(7)	-16.21(1)

^aValues for the electronic ground states from [37].^bRelative to the electronic origin of 7AI at 34630.74 cm⁻¹.**Figure 3.** Rotationally resolved LIF spectrum of 7AI, 7AI[ND], 7AI[CD] and 7AI[ND][CD]. The upper trace gives the experimental spectrum, the second trace the simulation, using the best fit parameters. The following traces show simulations of the individual spectra of 7AI, 7AI[ND], 7AI[CD] and 7AI[ND][CD]. They are given only for reason of clarity, the fit has been performed using the overall spectrum. The frequency scale is relative to the origin of 7AI. The electronic origins of the other isotopomers are marked by arrows.

and/or if only a partial spectrum is available the method is still successful. The GA succeeds in assigning the spectra and determines the molecular parameters without any prior knowledge of their values.

The success of the GA procedure of automated fitting is based on the existence of a good model for the prediction of the spectra. This seems to be the only drawback until now. However, there are many cases for which a good model prediction exists in particular in absorption, cavity ringdown and laser-induced fluorescence spectra. Based on our experience with the method, it is clear that in the case of small and/or local perturbations the main spectral features

that conform to the model can still be extracted and hence the perturbations are isolated.

The examples discussed above and references cited [16–21], demonstrate the extreme power of the GA in automated fitting and assigning very complex spectra, spectra which can hardly be analysed with the conventional methods. The computing power of modern PCs is more than adequate to perform the job in an acceptable time. This new technique opens the road to the analysis of the complex spectra of biomolecules and their building blocks. This has recently been demonstrated in a study of the biomolecule tryptamine and its complex with water [22].

Acknowledgments

We thank Jos Hageman, Ron Wehrens and Gerrit Groenenboom for many helpful discussions. The financial support of the Deutsche Forschungsgemeinschaft (SCHM 1043/9-4) is gratefully acknowledged. MS thanks the Nordrheinwestfälische Akademie der Wissenschaften for a grant which made this work possible. The authors thank the National Computer Facilities of the Netherlands Organization of Scientific Research (NWO) for a grant on the Dutch supercomputing facility SARA.

References

- [1] Majewski W A and Meerts W L 1984 *J. Mol. Spectrosc.* **104** 271–81
- [2] Loomis F W and Wood R W 1928 *Phys. Rev.* **32** 223–36
- [3] Scott J F and Rao K N 1966 *J. Mol. Spectrosc.* **20** 461–3
- [4] Winnewisser B P, Reinstädler J, Yamada K M T and Behrend J 1989 *J. Mol. Spectrosc.* **136** 12–6
- [5] Neese C F 2001 *Int. Symp. on Molecular Spectroscopy, 56th Meeting, 11–15 June*
- [6] Thompson C D, Robertson E G and McNaughton D 2003 *Phys. Chem. Chem. Phys.* **5** 1996–2000
- [7] Plusquellic D F, Suenram R D, Mate B, Jensen J O and Samuels A C 2001 *J. Chem. Phys.* **115** 3057–67
- [8] Moruzzi G 2005 *J. Mol. Spectrosc.* **229** 19–30
- [9] Helm R M, Vogel H-P and Neusser H J 1997 *Chem. Phys. Lett.* **270** 285–91
- [10] Haynam C A, Brumbaugh D V and Levy D H 1984 *J. Chem. Phys.* **81** 2282–94
- [11] Philips L A and Levy D H 1986 *J. Chem. Phys.* **85** 1327–32
- [12] Philips L A and Levy D H 1988 *J. Chem. Phys.* **89** 85–90
- [13] Hageman J A, Wehrens R, de Gelder R, Meerts W L and Buydens L M C 2000 *J. Chem. Phys.* **113** 7955–62
- [14] Berden G, Meerts W L and Jalviste E 1995 *J. Chem. Phys.* **103** 9596–606
- [15] Berden G, van Rooy J, Meerts W L and Zachariasse K A 1997 *Chem. Phys. Lett.* **278** 373–9
- [16] Szydłowska I, Myszkiewicz G and Meerts W L 2002 *Chem. Phys.* **283** 371–7
- [17] Schmitt M, Ratzer C and Meerts W L 2004 *J. Chem. Phys.* **120** 2752–8
- [18] Meerts W L, Schmitt M and Groenenboom G 2004 *Can. J. Chem.* **82** 804–19
- [19] Schmitt M, Ratzer C, Kleinermanns K and Meerts W L 2004 *Mol. Phys.* **102** 1605–14
- [20] Nikolaev A E, Myszkiewicz G, Berden G, Meerts W L, Pfanstiel J F and Pratt D W 2005 *J. Chem. Phys.* **122** 084309–1–10
- [21] Myszkiewicz G, Meerts W L, Ratzer C and Schmitt M 2005 *Phys. Chem. Chem. Phys.* **7** 2142–50
- [22] Schmitt M, Böhm M, Ratzer C, Vu C, Kalkman I and Meerts W L 2005 *J. Am. Chem. Soc.* **127** 10356–64
- [23] Metzger G J, Patel M and Hu X 1996 *J. Magn. Reson. B* **110** 316–20
- [24] Dods J, Gruner D and Brumer P 1996 *Chem. Phys. Lett.* **261** 612–9
- [25] Ahonen H, de Souza P A Jr and Vijayendra K G 1997 *Nucl. Instrum. Methods Phys. B* **124** 633–8
- [26] Golovkin I, Mancini R, Louis S, Lee R and Klein L 2002 *J. Quantum Spectrosc. Radiat. Transfer* **75** 625–36
- [27] Spalek T, Pietrzyk P and Sojka Z 2005 *J. Chem. Inf. Model.* **45** 18–29
- [28] Holland J H 1975 *Adaption in Natural and Artificial Systems* (Ann Arbor, MI: University of Michigan Press)
- [29] Goldberg D E 1989 *Genetic Algorithms in Search, Optimisation and Machine Learning* (Reading, MA: Addison-Wesley)
- [30] Rechenberg I 1973 *Evolutionsstrategie—Optimierung technischer Systeme nach Prinzipien der biologischen Evolution* (Stuttgart: Frommann-Holzboog)
- [31] Berden G, Meerts W L, Schmitt M and Kleinermanns K 1996 *J. Chem. Phys.* **104** 972–82
- [32] Schmitt M, Küpper J, Spangenberg D and Westphal A 2000 *Chem. Phys.* **254** 349–61
- [33] Allen H C and Cross P C 1963 *Molecular Vib-Rotors* (New York: Wiley)
- [34] Wu Y R and Levy D H 1989 *J. Chem. Phys.* **91** 5278–84
- [35] Levine D 1996 PGAPack V1.0, PgaPack can be obtained via anonymous ftp from <ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.z>
- [36] Borst D R, Korter T M and Pratt D W 2001 *Chem. Phys. Lett.* **350** 485–90
- [37] Caminati W, di Bernardo S and Trombetti A 1990 *J. Mol. Struct.* **223** 415–24