New applications of the genetic algorithm for the interpretation of high-resolution spectra[1]

W. Leo Meerts, Michael Schmitt, and Gerrit C. Groenenboom

Abstract: Rotationally resolved electronic spectroscopy yields a wealth of information on molecular structures in different electronic states. Unfortunately, for large molecules the spectra get rapidly very congested owing to close-lying vibronic bands, other isotopomers with similar zero-point energy shifts, or large-amplitude internal motions. A straightforward assignment of single rovibronic lines and, therefore, line position assigned fits are impossible. An alternative approach is unassigned fits of the spectra using genetic algorithms (GAs) with special cost functions for evaluation of the quality of the fit. This paper decribes the improvements we established on the GA method discussed before (J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, and L.M.C. Buydens. J. Chem. Phys. **113**, 7955 (2000)). In particular, we succeeded in obtaining a dramatic reduction in computing time that made it possible to apply the GA process in a large number of cases. A completely automated fit of a rotationally resolved laser-induced fluorescence spectrum without any prior knowledge of the molecular parameters can now be performed in less than 1 h. We demonstrate the power of the method on a number of typical examples such as very dense rovibronic spectra of van der Waals clusters and overlapping spectra due to different isotopomers. The discussed results demonstrate the extreme power of the GA in automated fitting and assigning of complex spectra. It opens the road to the analysis of complex spectra of biomolecules and their building blocks.

Key words: high-resolution spectroscopy, genetic algorithm, biomolecules, structure, van der Waals clusters.

Résumé : La spectroscopie électronique résolue en fonction de la rotation fournit une profusion d'informations relatives aux structures moléculaires dans les différents états électroniques. Malheureusement, dans les cas de molécules de taille importante, les spectres deviennent rapidement très congestionnés en raison des bandes vibroniques adjacentes, d'autres isotopomères ayant des déplacements semblables de l'énergie du point zéro ou des mouvements internes de grandes amplitudes. Il est toutefois impossible de faire des attributions non ambiguës des raies rovibroniques simples et, par extension, des ajustements des positions de raies attribuées. Une autre méthode est de faire l'ajustement des raies non attribuées des spectres en faisant appel à des algorithmes génétiques (AG) et en appliquant des fonctions spéciales de coûts pour l'évaluation de la qualité des ajustements. Dans ce travail, on décrit les améliorations qu'on a apporté à la méthode des AG discutée antérieurement (J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, et L.M.C. Buydens. J. Chem. Phys. **113**, 7955 (2000)). On a réussi, en particulier, à obtenir une diminution dramatique du temps de calcul nécessaire pour appliquer le processus d'AG à un grand nombre de cas. Il est maintenant possible de réaliser en moins d'une heure un ajustement complètement automatique d'un spectre de fluorescence induite au laser et résolue en fonction de la rotation, sans avoir recours à une connaissance préalable des paramètres moléculaires. On a démontré la puissance de la méthode sur un grand nombre d'exemples typiques, tels que les spectres rovibroniques très denses d'agrégats de van der Waals et les spectres comportant des recouvrements en raison d'isotopomères différents. Les résultats discutés démontrent la très grande puissance des AG dans l'ajustement automatique et l'attribution de spectres complexes. Ils ouvrent la voie à une analyse des spectres complexes de biomolécules et de leur composantes.

Mots clés : spectroscopie à haute résolution, algorithme génétique, biomolécules, structure, agrégats de van der Waals.

[Traduit par la Rédaction]

1. Introduction

Rotationally resolved electronic spectroscopy provides a valuable tool for determination of molecular structures in different electronic states. An implicit problem of this method is that for larger molecules the spectra rapidly get very congested. Additionally, overlapping bands due to (*i*) close-spaced vibronic bands, (*ii*) other isotopomers with similar zero-point energy

Received 3 November 2003. Published on the NRC Research Press Web site at http://canjchem.nrc.ca/ on 26 August 2004.

W.L. Meerts.[2] Department of Molecular and Laser Physics, NSRIM, University of Nijmegen, P.O. Box 9010, NL-6500 GL Nijmegen, The Netherlands.
M. Schmitt. Heinrich-Heine-Universität, Institut für Physikalische Chemie, D-40225 Düsseldorf, Germany.
G.C. Groenenboom. Institute of Theoretical Chemistry, NSRIM, University of Nijmegen, P.O. Box 9010, NL-6500 GL Nijmegen, The Netherlands (e-mail: gerritg@theochem.kun.nl).

[1]This article is part of a Special Issue dedicated to the memory of Professor Gerhard Herzberg.
[2]Corresponding author (e-mail: Leo.Meerts@sci.kun.nl).

 doi: 10.1139/V04-037

shifts, or (*iii*) split bands due to large-amplitude internal motions might complicate the experimental spectrum further. All these factors make a straightforward assignment of single rovibronic lines and, therefore, line position assigned fits impossible. In Neusser's group (1), a procedure has been developed that directly fits the experimental data without prior assignments. This method, which is called "correlation automated rotational fitting", was pioneered by Levy and co-workers (2–4) 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. An alternative approach is unassigned fits of the spectra using genetic algorithms (GAs) with special cost functions for evaluation of the level of the fit.

It has been shown by Hageman et al. (5) that a GA with a properly defined cost function was capable of performing automated fitting of spectra without any prior knowledge of the molecular parameters. The cost function used by Hageman et al. (5) is able to smooth the error landscape and, therefore, allows the GA to locate the global minimum. Unfortunately, this method is quite time-consuming, compared to other cost functions like simple least-squares or peak picking functions. The automated fitting of several overlapping bands requires, therefore, fast parallel processing and long computing times. In the present paper, we show how the computing time of the cost function can be reduced drastically, so that the automated fit of a rovibronic spectrum can be performed in less than 1 h using a standard desktop PC. The performance of the GA for spectral simulations has been described in detail elsewhere (5). A good introduction to the vocabulary and theory of GAs as a tool for solving optimization problems can be found in refs. 6 and 7.

In this paper, we extend the automated fit to the case of several overlapping spectra, i.e., the fitting of molecular parameters that belong to different molecular species or spectral components. The method is applied to a synthetic spectrum, which consists of two completely overlapping bands, to adapt the internal parameters for the GA fit. The refined method is then applied to a series of experimental rovibronic spectra of isotopomers of phenol and benzonitrile, and clusters thereof. The discussed results demonstrate the extreme power of the GA in automated fitting and assigning of complex spectra.

2. Experimental

The experimental setup for the rotational resolved laser-induced fluorescence (LIF) is described elsewhere (8). Briefly, it consists of a ring dye laser (Coherent 899-21) operated with Rhodamine 110, pumped with 6 W of the 514 nm line of an Ar^+-ion laser. The light is coupled into an external folded ring cavity (9) for second-harmonic generation (SHG). The molecular beam is formed by expanding the vaporized sample, seeded in 400−1000 mbar of argon (1 bar = 100 kPa), through a 70 μm hole into the vacuum. The molecular beam machine consists of three differentially pumped vacuum chambers that are linearly connected by skimmers (1 and 3 mm, respectively) to reduce the Doppler width. The molecular beam is crossed at right angles in the third chamber with the laser beam 360 mm downstream of the nozzle. The resulting fluorescence is collected perpendicular to the plane defined by the laser and molecular beam by an imaging optics setup consisting of a concave mirror and two plano-convex lenses. The resulting Doppler width in this setup is 25 MHz (fwhm). The integrated molecular fluorescence is detected by a photomultiplier tube, and the output is discriminated and digitized by a photon counter and transmitted to a PC for data recording and processing. The relative frequency is determined with a quasi-confocal Fabry−Perot interferometer with a free spectral range (FSR) of 149.9434(56) MHz. The FSR was calibrated using the combination differences of 111 transitions of indole for which the microwave transitions are known (10, 11). The absolute frequency was determined by recording the iodine absorption spectrum and comparing the transitions to the tabulated lines (12).

3. Theory

3.1. The Hamiltonian

For the simulation of the rovibronic spectra, a rigid asymmetric rotor Hamiltonian was employed (13):

$$[1] \qquad \widehat{H} = A P_a^2 + B P_b^2 + C P_c^2$$

Here, $P_g (g = a, b, c)$ are the components of the body-fixed angular momentum operator, and A, B, and C are the three rotational constants. The resulting Hamiltonian matrix is factorized into four submatrices using the Wang transformation (14). This enhances the computation speed because of the reduced dimension of the matrix to be diagonalized. The transition frequencies are determined by the rotational constants A, B, C in both electronic states and by the frequency of the origin ν_0 of the vibronic band. The line intensities depend on the rotational temperature, the orientation of the transition dipole moment with respect to the inertial axes, and, in some cases, the nuclear spin statistic. The temperature dependence of the intensities might be considered in a simple one-temperature model, which is sufficient for simulation of most of the rovibronic spectra, or in a more advanced two-temperature model, which should be applied in cases where line shape parameters are fitted (cf. Sect. 4.4.2). The orientation of the dipole moment vector is determined by the polar angle θ and the azimuthal angle ϕ.

3.2. The genetic algorithm

A description of the GA used in this investigation can be found in ref. 5. The GA library PGAPack version 1.0, which can run on parallel processors, was used (15). The calculations were performed on four processors of a SUN UltraSPARC 333 MHz and on a 2.6 GHz PC with two processors under Linux. 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 (16–18). We introduce the elements of the GA that 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 of the quality of the individual solution. The fitness function that is used here is described in Sect. 3.3.
- 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 and 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 probability, 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 Sect. 4.2.

3.3. The fitness function

3.3.1. Definition

In the current paper we will use both the terms *fitness function* (F_{fg}) and *cost function* (C_{fg}), where $C_{fg} = 100(1 - F_{fg})$, to characterize the quality of a solution. The fitness function for the fit of the spectra with N points using the GA has been defined in eq. [5] of ref. 5 (in which C_{fg} is identical to F_{fg} in this paper) as:

$$[2] \qquad F_{fg} = \frac{\sum_{r=-l}^{l} w(r) \sum_{i=1}^{N} f(i)g(i+r)}{\sqrt{\sum_{r=-l}^{l} w(r) \sum_{i=1}^{N} f(i)f(i+r)}\sqrt{\sum_{r=-l}^{l} w(r) \sum_{i=1}^{N} g(i)g(i+r)}}$$

In this equation, f and g represent the experimental and calculated spectra, respectively. The function $w(r)$ determines the sensitivity of the fitness function for a shift of the two spectra relative to each other. This can be rewritten by interchanging the sums and substituting $i + r = j$ as

$$[3] \qquad F_{fg} = \frac{\sum_{i,j}^{N} f_i W_{ij} g_j}{\sqrt{\sum_{i,j}^{N} f_i W_{ij} f_j}\sqrt{\sum_{i,j}^{N} g_i W_{ij} g_j}}$$

where

$$[4] \qquad W_{ij} = w(|j - i|)$$

and $f_i = f(i)$ and $g_i = g(i)$.

F_{fg} in eq. [3] can be interpreted as the cosine of the "angle" between the experimental and the theoretical spectrum. With the column vectors

$$[5] \qquad \boldsymbol{f} = (f_1, f_2, ..f_N)^T$$
$$\boldsymbol{g} = (g_1, g_2, ..g_N)^T$$

and the symmetric matrix $\boldsymbol{W}$, which has the matrix elements W_{ij}, we can write:

$$[6] \qquad F_{fg} = \cos(\alpha) = \frac{(\boldsymbol{f}, \boldsymbol{g})}{\|\boldsymbol{f}\| \, \|\boldsymbol{g}\|}$$

Here the inner product $(\boldsymbol{f}, \boldsymbol{g})$ is defined with the metric $\boldsymbol{W}$ as:

$$[7] \qquad (\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{f}^T \boldsymbol{W} \boldsymbol{g}$$

and the norm of $\boldsymbol{f}$ as $\|\boldsymbol{f}\| = \sqrt{(\boldsymbol{f}, \boldsymbol{f})}$; similarly for $\boldsymbol{g}$. For $w(r)$ we used a triangle function (5) with a width the base of Δw:

$$[8] \qquad w(r) = \begin{cases} 1 - |r| \,/\, (1/2\Delta w) & \text{for } |r| \leqslant 1/2\Delta w \\ 0 & \text{otherwise} \end{cases}$$

In order for F_{fg} to serve as a good fitness function for the quality of the fit, it should have the property that it reaches its maximum value if and only if $\boldsymbol{f}$ and $\boldsymbol{g}$ are identical (apart from a normalization). This condition is fulfilled provided the matrix $\boldsymbol{W}$ is positive definite. In Appendix A, we show this holds if the Fourier transform of the function w is positive. The Fourier transform of $w(r)$ defined in eq. [8] is

$$[9] \qquad \tilde{w}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{-2\pi i r t} w(r) \, \mathrm{d}r = \frac{\Delta w}{2} \, \mathrm{sinc}^2 \left(\frac{\Delta w}{2} \pi t \right)$$

where $\mathrm{sinc}\,(x) = \sin(x)/x$, so $\tilde{w}(t) \geqslant 0$.

3.3.2. Numerical evaluation of the fitness

Let us now consider the numerical evaluation of the fitness function F_{fg} from eq. [3] in its relation with the calculated spectrum. The calculated spectrum is obtained by a convolution of each calculated transition k with intensity s_k by the line shape function L:

$$[10] \qquad g_j = \sum_{k=1}^{N} L_{jk} s_k$$

with $L_{jk} = l(|j - k|)$. In matrix notation this can be rewritten as

[11] $$\boldsymbol{g} = \boldsymbol{L}\boldsymbol{s}$$

As it turned out, at least 50% of the computing time in ref. 5 was used to perform a straightforward calculation of F_{fg} from eq. [3]. For a typical GA fit, F_{fg} must be calculated 150 000 times. Hence, a considerable reduction in computing time can be obtained by a more efficient calculation of F_{fg}. We start with a rearrangement of the order of the evaluation of eq. [6] and using the properties of $\boldsymbol{W}$ and $\boldsymbol{L}$. The numerator of eq. [6] is evaluated first:

[12] $$(\boldsymbol{f}, \boldsymbol{L}\boldsymbol{s}) = \boldsymbol{f}^T\boldsymbol{W}\boldsymbol{L}\boldsymbol{s} = \tilde{\boldsymbol{f}}\boldsymbol{s}$$

[13] $$\tilde{\boldsymbol{f}} = \boldsymbol{f}^T\boldsymbol{W}\boldsymbol{L}$$

From eqs. [12] and [13] it is obvious that the effect of $w(r)$ can be interpreted as an effective line broadening of the experimental[3] spectrum. The use of the broadening function $w(r)$ results in a smoother error landscape, which allows an easier optimization of the GA process. It should be noted that the simple least-squares fitness function, where all spectral points have the same weight, is also described by F_{fg} for the limiting case that the width of $w(r)$ is zero and $\boldsymbol{W}$ becomes the identity matrix.

The transformed experimental spectrum from eq. [13] has to be evaluated only once. Formally, the sum on the right-hand side of eq. [12] runs over all N points of the spectrum. However, the stick spectrum array is a very sparse one: Typically, N is of the order of 60 000 − 250 000, while the number of sticks (nonzero values of s_k) is only about 1000−3000. Therefore, the use of eq. [12] strongly reduces the necessary computing time. The reduction of computing time is actually more dramatic since the double sum in the numerator of eq. [3] over N points is reduced to a single sum over a sparse array in eq. [12]. Furthermore, the theoretical spectrum itself $\{g_i\}$ does not have to be calculated anymore. The first term in the denominator of eq. [3] also has to be evaluated only once. The second term in this denominator has to be calculated every time a value of F_{fg} is needed in the process of the GA. Fortunately, this term can also be expressed in terms of the stick spectrum $\boldsymbol{s}(\{s_k\})$:

[14] $$\|\boldsymbol{g}\|^2 = \boldsymbol{s}^\dagger(\boldsymbol{L}^\dagger\boldsymbol{W}\boldsymbol{L})\boldsymbol{s}$$

Since $(\boldsymbol{L}^\dagger\boldsymbol{W}\boldsymbol{L})$ is a banded matrix, the evaluation of $\|\boldsymbol{g}\|^2$ from eq. [14] is in practice almost linear in the number of sticks rather than quadratic. Again, $(\boldsymbol{L}^\dagger\boldsymbol{W}\boldsymbol{L})$ has to be evaluated only once. It turned out that the effect of the above-discussed modifications of the calculation of F_{fg} was that its calculation time became negligible with respect to the total computing time.

The use of the stick spectrum, described in this section, is limited to applications in which the line shape parameters, like Gaussian or Lorentzian width in the Voigt profile, do not have to be fitted. Inclusion of the line width parameters in the fit requires the reevaluation of eq. [12] during the GA process, which in practice dramatically increases the computing time. Therefore, a line width fit should be performed after a determination of all other parameters in a separate fitting procedure. An example of this will be given in Sect. 4.4.2.

3.3.3. Further reduction of computing time

Further reduction of the computing time for the fitness function is made possible by setting the maximum J value in the evaluation of the simulated spectrum dynamically. In the first step of the evaluation, a maximum J value is taken, which is specified by the user. In the subsequent steps, the necessary $J_{\max}$ is computed from the temperature and the rotational constants using a cut-off factor of 0.001 for the intensities. In this way, the size of the matrices to be set up and diagonalized in the course of the computation of the simulated spectrum is minimized.

4. Results and discussion

4.1. Influence of the width (Δw) of the weight function *w(r)* on the convergence of the GA

The relative broadening of the spectra, introduced by the weight function $w(r)$, described in Sect. 3.3.2, critically determines the ability of the GA to converge to the global minimum and also the speed of convergence. The smoothing of the error landscape allows the sensing of regions far from the minimum. In the first step, the function $w(r)$ should be chosen relatively broad; $\Delta w \approx$ 15–20 times the line widths of an individual rovibronic line in the spectrum (Δ_{lw}). In this way, a first set of molecular parameters is obtained, which still has to be refined. This is done by decreasing Δw and narrowing the limits of the parameter space to be searched in the fit. Decreasing Δw improves the accuracy of the molecular parameters obtained from the fit, while narrowing the parameter space leads to an improved sampling in the region of the minimum. This of course is the critical step in the procedure. Too strong narrowing of the parameter space leads to premature convergence of the fit — with a high probability into a local minimum.

We performed a fit of a synthetic spectrum that consisted of two overlapping sub-spectra. It was generated using the molecular parameters from the "Best value" column in Table 1. A single-temperature model for the calculation of the intensities has been used. The maximum J value used in the computation of the cost function for this spectrum is 22. Table 2 lists the results of five GA calculations stopped after 500 generations, each started with a different randomly generated initial set of parameters. In Table 2, $\Delta w/\Delta_{lw} = 20$. This process has been repeated for different values of $\Delta w/\Delta_{lw}$. The convergence of the fit using different $\Delta w/\Delta_{lw}$ is shown in Fig. 1.

A ratio of $\Delta w/\Delta_{lw} = 20$ leads to convergence for all five initial seeds into the same minimum. Inspection of the parameters in Table 2 shows that the minimum found is the global one. Nevertheless, the deviations of the fitted parameters from the best values of Table 1 are quite large, owing to the large value of Δw, which leads to broad minima at the cost surface. As seen in Fig. 1, smaller values of Δw may lead to convergence into other (local) minima of the cost surface. With $\Delta w/\Delta_{lw}$ of 10, still three of the fits converge to the global minimum, with a ratio of 5, only one fit converges to the global minimum, and with 2.5, none of the fits converges to the global minimum.

[3] A different arrangement of the equations shows that $w(r)$ also can be interpreted as an effective broadening of the calculated spectum g. Actually, the experimental and the calculated spectra are broadened relative to each other.

Table 1. Input for the genetic algorithm (GA) fit of the synthetic spectrum.

Parameter	Best value	Lower limit	Upper limit	Coupling [a]
Parameters of the first spectrum				
A_1''	3 000.00	2 900.00	3 100.00	
B_1''	1 000.00	910.00	1 100.00	
C_1''	800.00	700.00	900.00	
T_1	2.00			
θ_1	54.74	0.00	90.00	
ϕ_2	45.00	0.00	90.00	
$\nu_0^{(1)}$	80 000.00	79 000.00	81 000.00	
ΔA_1	100.00	10.00	150.00	
ΔB_1	−50.00	−150.00	50.00	
ΔC_1	−50.00	−150.00	50.00	
Gaussian width	20.00			
Parameters of the second spectrum				
$A_2'' - A_1''$	10.00	5.00	15.00	D
$B_2'' - B_1''$	5.00	0.00	10.00	D
$C_2'' - C_1''$	−5.00	−10.00	0.00	D
T_1	2.00			C
θ_2	56.74			
ϕ_2	44.00			
$\Delta\nu_0^{(2)} - \Delta\nu_0^{(1)}$	500.00	0.00	1 000.00	D
$\Delta\Delta A$	2.00	−3.00	3.00	D
$\Delta\Delta B$	−1.00	−3.00	3.00	D
$\Delta\Delta C$	1.00	−3.00	3.00	D
Gaussian width	20.00		C	
Ratio scaling	1.00		R	

Note: All values are in MHz except T, which is in K, and θ and ϕ, which are in degrees. The Lorentzian contribution to the line width was sert to zero. $\Delta A_1 = A_1' - A_1''$; ΔB_1 and ΔC_1 correspondingly. $\Delta\Delta A = \Delta A_2 - \Delta A_1 = \left(A_2' - A_2''\right)\left(A_1' - A_1''\right)$; etc.

[a] C, parameters of spectrum 2 are taken the same as those of spectrum 1; D, parameters of spectrum 2 are those of spectrum 1 with the corresponding value in the table added; R, parameters of spectrum 2 are those of spectrum 1 multiplied by the corresponding value from the table.

In the next step, the fit has to be refined with a limited parameter space centered around the best fit (fit No. 4 in Table 2) and with successively smaller values of Δw. We could not establish a hard criterion for the reduction of the parameter space. A successfull reduction depends critically on the quality of the first series of fits. On the other hand, the parameter space cannot always be reduced by the same factor for each of the parameters. As a rule of thumb, we found that the parameter limits can be reduced to one-tenth of the initial range. This reduction depends on the quality (signal-to-noise (S/N)) of the spectrum and has to be checked carefully after each successive step. Table 3 gives the new input parameters for a refined fit, using a ratio $\Delta w/\Delta_{lw} = 5$, along with the result of the best of five fits with different starting populations.

As can be inferred from Table 3, the fit with the reduced line width ratio and the reduced parameter space is already quite close to the "real" values given in Table 1. It can further be improved by decreasing $\Delta w/\Delta_{lw}$ and the parameter search space in an iterative manner until convergence for the molecular parameters is reached.

To generate a more realistic spectrum, we added a randomly Gaussian-distributed noise to the synthetic spectrum, resulting in a S/N of 10:1 for the strongest lines. The GA performed equally well in this case, yielding the molecular parameters given in the last column of Table 3. The deviations of the parameters from the true values are similarly small as for the "perfect" spectrum without noise.

4.2. Meta-optimization of the internal GA parameters

The need to optimize the internal parameters of the GA (meta-optimization) for a given problem has been discussed to be a major drawback of this method (19). The speed and convergence of GAs depend on the data representation (binary or real), the crossover type (one-point, two-point, uniform), the size of the starting population, the rate of elitism, and the mutation probability. Several other factors that also influence the performance of the GA fit have been kept fixed. They will be discussed shortly. The crossover probability was chosen as 85%. Selection of the best solutions is performed via a tournament selection, which means that a random subset of the chromosomes is taken, and within each subset the chromosomes are selected by their cost. Duplicates within one generation are allowed for.

Table 2. Results of five successive GA evaluations of the synthetic spectrum from Table 1.

Parameter	Best value	Fit No. 1	Fit No. 2	Fit No. 3	Fit No. 4	Fit No. 5
Parameters of the first spectrum						
A_1''	3 000.00	2 996.19	2 998.14	2 998.53	3 000.68	2 996.38
B_1''	1 000.00	999.71	998.14	999.12	998.92	999.51
C_1''	800.00	801.49	799.90	797.95	800.88	800.68
θ_1	54.74	56.07	53.09	53.38	54.80	57.68
ϕ_2	45.00	45.17	44.68	45.42	45.71	42.29
$\nu_0^{(1)}$	8 0000.00	79 987.29	79 981.43	79 985.34	80 002.93	79 997.07
ΔA_1	100.00	101.61	101.81	101.42	99.46	101.81
ΔB_1	−50.00	−49.36	−48.58	−50.34	−49.17	−49.85
ΔC_1	−50.00	−49.95	−50.05	−49.27	−50.64	−50.34
Parameters of the second spectrum						
$A_2'' - A_1''$	10.00	13.53	11.22	11.93	9.52	15.07
$B_2'' - B_1''$	5.00	6.42	7.03	5.87	5.68	5.52
$C_2'' - C_1''$	−5.00	−5.31	−5.99	−2.45	−6.34	−6.29
θ_2	56.74	55.73	58.32	56.95	57.39	59.54
ϕ_2	44.00	47.03	42.68	43.56	44.34	42.09
$\Delta\nu_0^{(2)} - \Delta\nu_0^{(1)}$	500.00	505.38	520.04	509.29	488.76	501.47
$\Delta\Delta A$	2.00	0.49	0.71	0.54	2.92	−1.09
$\Delta\Delta B$	−1.00	−1.72	−2.44	−0.19	−2.00	−0.09
$\Delta\Delta C$	1.00	1.23	1.02	0.23	2.13	1.17
C_{fg}	0.000	0.394	0.526	0.397	0.279	0.571

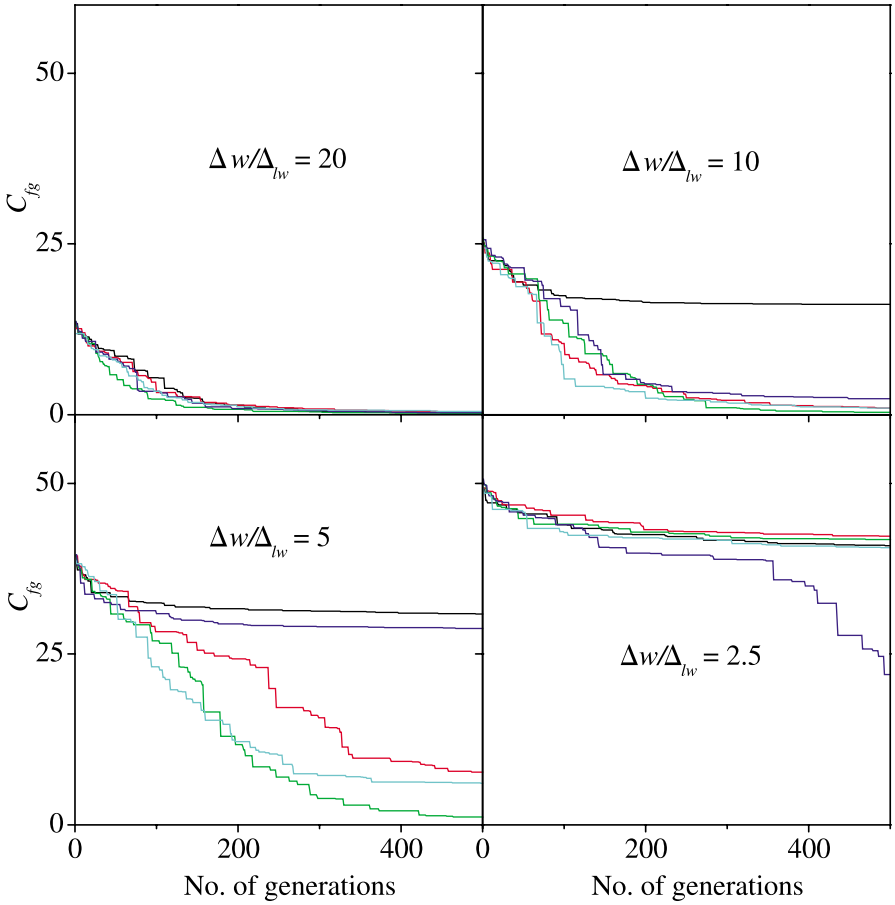
Note: The starting set of parameters was generated randomly and the search limits are from Table 1, $\Delta w/\Delta_{lw} = 20$.

Table 3. Results of a GA evaluation of the synthetic spectrum of Table 1 with narrowed search regions and $\Delta w/\Delta_{lw} = 5$.

Parameter	Best fit	Lower limit	Upper limit	Best fit with noise
Parameters of the first spectrum				
A_1''	3 000.19	2 990.00	3 010.00	3 000.01
B_1''	1 000.22	990.00	1 010.00	1 000.01
C_1''	800.05	790.00	810.00	799.76
θ_1	55.11	40.00	60.00	53.74
ϕ_2	44.63	40.00	60.00	45.47
$\nu_0^{(1)}$	80 000.05	79 950.00	80 150.00	80 000.44
ΔA_1	100.11	90.00	110.00	99.91
ΔB_1	−50.12	−60.00	−30.00	−49.88
ΔC_1	−49.94	−60.00	−30.00	−50.12
Parameters of the second spectrum				
$A_2'' - A_1''$	9.66	5.00	15.00	10.44
$B_2'' - B_1''$	4.59	0.00	10.00	5.18
$C_2'' - C_1''$	−5.15	−10.00	0.00	−4.80
θ_2	56.99	40.00	60.00	56.30
ϕ_2	45.08	40.00	60.00	43.17
$\Delta\nu_0^{(2)} - \Delta\nu_0^{(1)}$	497.17	400.00	600.00	501.66
$\Delta\Delta A$	2.08	0.00	4.00	1.78
$\Delta\Delta B$	−0.85	−4.00	0.00	−1.17
$\Delta\Delta C$	0.97	0.00	4.00	1.15
C_{fg}	0.162	—	—	0.297

The "natural" choice for a genetic algorithm is a binary representation of the data. It has been discussed that a direct representation of the parameters as real-type data is advantageous if the type of data to be fitted is real. All tests performed with the data set given in Table 1 show that the binary encoding of the parameters leads to a smaller cost and converges more rapidly, compared to real-type representation. For the real-type data encoding, a number of runs do not even reach the global minimum.

Fig. 1. Dependence of the convergence of five fits of a synthetic spectrum that consists of two overlapping sub-spectra with different starting populations on the ratio $\Delta w/\Delta_{lw}$.



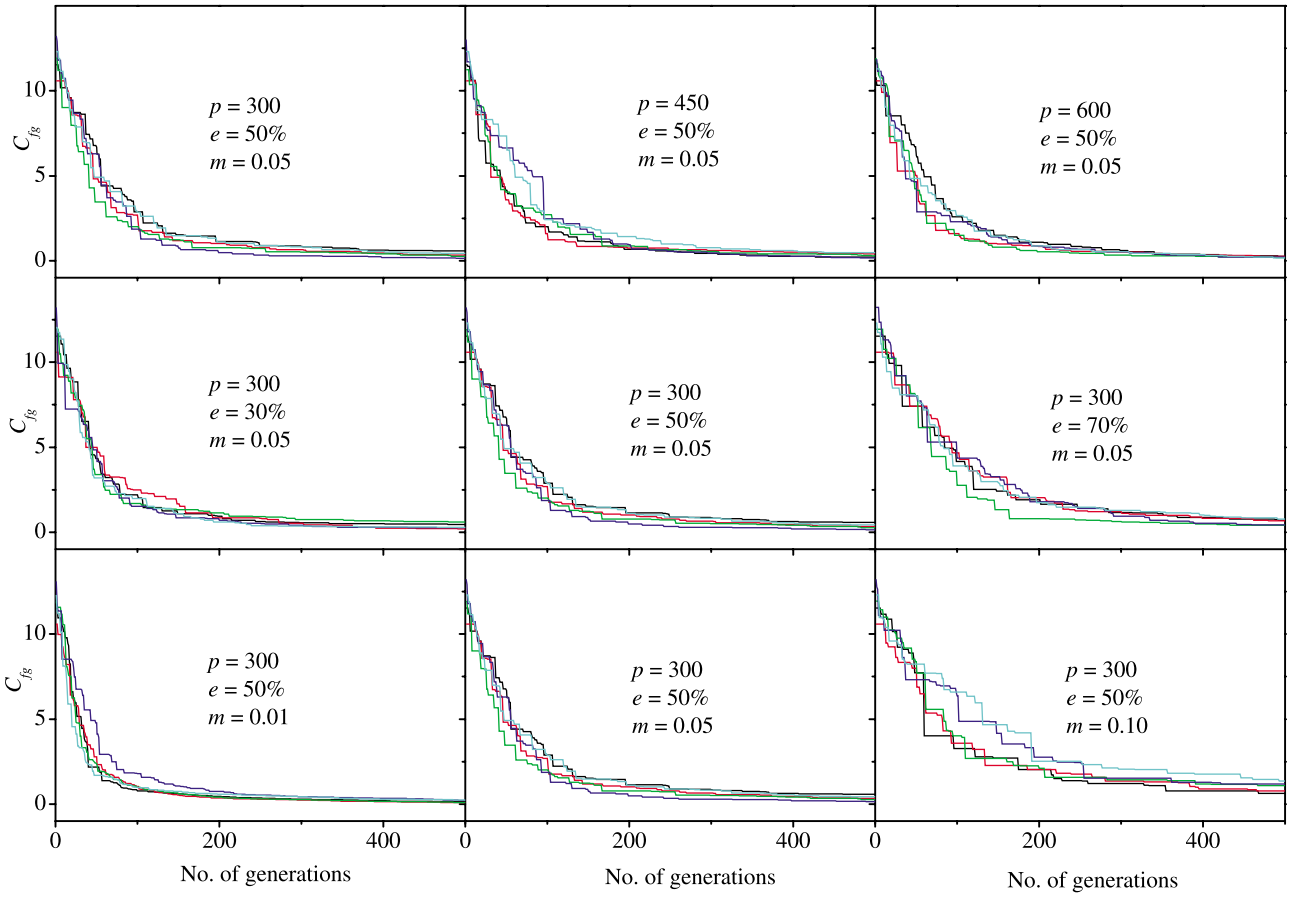
A change of the encryption depth for the binary representation from 10 bit to 20 bit virtually does not change the performance of the GA. A Gray code (20) is used throughout the present investigations in order to ensure a Hamming distance of one (21). Comparing uniform and two-point crossover, a clear advantage of the two-point crossover regarding speed of convergence and fitness of the solutions was found. Just in the case of real-type encoding, the uniform crossover forces more runs of the GA into the global minimum than the two-point.

The first row of Fig. 2 shows the convergence of five fits with different starting populations as a function of the number of generations for different sizes of the starting population, using the parameters from Table 1 and $\Delta w/\Delta_{lw} = 20$ (cf. Sect. 4.1). The elitism was kept at 50% and mutation probability at 0.05 in these calculations. For a population of 300, the best value of the cost function was 0.15, the mean of five runs using different starting populations was 0.35, and the cost function dropped below 0.5 after 372 generations. The larger population of 450 had a slightly better mean of 0.30 and dropped below 0.5 after 324 generations. The largest population we investigated contained 600 chromosomes and resulted in a mean of the cost function of 0.21 and dropped below 0.5 after 278 generations. The better performance regarding the convergence as a function of the number of generations for the larger populations is more than compensated for by the increasing CPU time for increasing populations. One run for a population of 300 takes 21 min, for a population of 450 takes 43 min, and for 600, 52 min using four processors on a SUN UltraSPARC 333 MHz. All subsequent computing times are for this configuration. Almost the same computing time was attained on a dual processor PC with two Pentium 2.6 GHz processors.

The variation of elitism between 30% and 70% is shown in the second row of Fig. 2. The size of the population is kept fixed at 300 for these fits. An elitism of 30% means that the best 30% of one generation are passed unchanged to the next generation. Elitism helps to prevent good solutions from being lost from one generation to the next. As can be inferred from Fig. 2, a fit with an elitism of 30% converges more rapidly than one with 50% or 70%. Nevertheless, the mean cost function of five runs for an elitism of 50% is slightly better. For an elitism of 70%, the mean cost function never drops below 0.5 because too many bad solutions are kept. With regard to CPU time, an elitism of 30% is the most time-consuming (36 min) compared to 50% (21 min) and 70% (12 min).

The third row of Fig. 2 shows the variation of convergence with the mutation probability. Population size is kept fixed at 300 and elitism at 50%. Mutation is introduced to allow for a thorough exploration of the whole cost surface and prevents the algorithm from prematurely converging into a local minimum. With a mutation probability of 0.01, the mean value of the cost function is 0.15, and the cost function drops below 0.5 after 204 generations. With a mutation probability of 0.05, the

Fig. 2. Convergence of the GA as a function of population size (p), rate of elitism (e), and mutation probability (m). The cost function C_{fg} is plotted on the vertical axis.



mean of the cost function increases to 0.35 (drops below 0.5 after 372 generations), while for a mutation probability of 0.10, the mean cost function value is 1.00. Although all five runs converged to the same global minimum for mutation probabilities of 0.01, 0.05, and 0.10, the mutation probability of 0.01 performed best, considering mean cost function and speed of convergence. However, one has to be careful not to underestimate the risk of converging to a local minimum due to bad exploration of the cost surface. The CPU time as a function of the mutation probability virtually does not change (21 min for probabilities of 0.01, 0.05, and 0.10).

To conclude, the population size should not exceed 300 because of the bad time performance with larger populations, which is not compensated by a much better convergence of the algorithm. An elitism of 30% is advantageous regarding the convergence, but also very time-consuming. The best trade-off between time and convergence performance is found at 50% elitism. A mutation probability of only 0.01 leads to a very fast convergence, with very exactly determined parameters. Nevertheless, in cases where many local minima at the cost surface are present, such a low mutation probability might lead into a local minimum. Reduction of the value of this parameter has therefore to be performed with great care.

4.3. GA fit of very dense rovibronic spectra

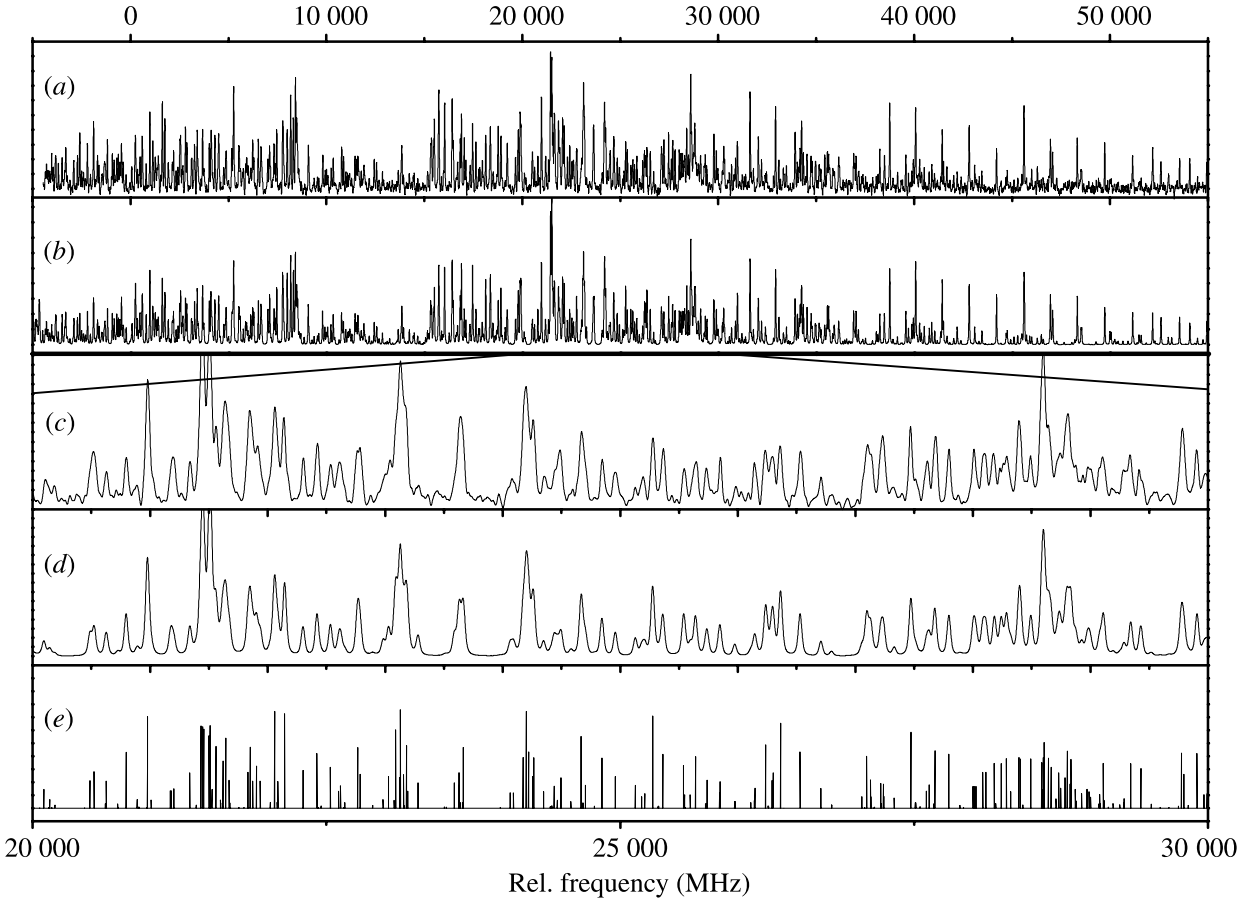
In the following, we will present the automated-GA fits of some rovibronic spectra, which are very congested due to small rotational constants. These spectra normally do not represent a great difficulty for the GA, as will be shown in the next sections.

4.3.1. [7-D]Phenol–N_2

We recently performed a fit of the rovibronic spectra of several isotopomers of the phenol–nitrogen cluster (22). The nitrogen is located in the plane of the phenol, hydrogen bonded to the OH group with a bond length of 225.5 pm. In the following, we present the rovibronic spectra of different [7-D]phenol clusters. [7-D]Phenol means replacement of hydrogen by deuterium at the hydroxy group of phenol. The choice of [7-D]phenol instead of the normal isotopomer was made because of the longer lifetime of the deuterated isotopomers, yielding smaller line widths and, therefore, better signal-to-noise ratios. The rotationally resolved electronic spectrum of the electronic origin of [7-D]phenol–N_2 is shown in trace (*a*) of Fig. 3. The observed spectrum consists of about 400 clusters of lines, with only a few single rovibronic lines (cf. the simulated stick spectrum shown in trace (*e*) of Fig. 3).

An assigned fit for such a congested spectrum is very dif-

Fig. 3. (*a*) Experimental rotationally resolved electronic spectrum of the electronic origin of phenol−N_2. (*b*) Simulation, using the parameters given in Table 4. (*c*) Expanded view of trace (*a*). (*d*) Simulation in the same spectral range with Voigt-convoluted line shapes, using a Gaussian width of 26 MHz and a Lorentzian width of 39.4 MHz. (*e*) Stick spectrum in the given spectral range. Intensities are given in arbitrary units.



ficult because the shape changes considerably upon moderate changes of the molecular parameters. The initial search range for the parameters in the GA fit was obtained from a preliminary ab initio calculation. This calculation was based on a "hydrogen" structure as proposed by Ford et al. (23). The molecular parameters obtained from the GA fit with $\Delta w/\Delta_{lw} = 5$ are presented in the second column of Table 4. The values given and the quoted uncertainties are the result of statistics on 10 independent GA runs, with different initial seeds, i.e., different starting populations of the evolution. In a second step, we used the result of the GA calculation to assign quantum numbers to the individual transitions and clusters of lines. Because of the high quality of the GA fit, this was possible in spite of the large number of overlapping lines. With these line position assignments, a second fit to the parameters of the rigid rotor Hamiltonian of eq. [1] was performed. The latter fit yields better values, in particular for the uncertainties of the parameters. For most parameters, the values obtained from the assigned fit (Table 4, column 3) agree within their uncertainties to the corresponding GA results. This spectrum presents an example in which the rovibronic spectrum could be fitted by the GA in a single step, without further refinement of Δw.

Table 4. Comparison of the molecular parameters of the phenol−nitrogen cluster as obtained from an assigned fit and from the GA fit.

Parameter	GA fit	Assigned fit
A'' (MHz)	4071.06 (13)	4072.18(25)
B'' (MHz)	647.89(2)	648.01(4)
C'' (MHz)	559.13(2)	559.26(4)
T (K)	1.6(5)	2.0
θ (°)	62.53(7)	60.0
ΔA (MHz)	−140.99(6)	−141.560(91)
ΔB (MHz)	15.71(1)	15.708(13)
ΔC (MHz)	8.70 (1)	8.671(9)

The computation time for the GA fit of the spectrum with 12 parameters and the direct evaluation of F_{fg} by the full sum of all data points (eq. [3]) was 12 min. It could be reduced to 5 min using the sparse stick array. Thus, a very complex spectrum could be completely fit by the GA within 50 min of computation time.

Table 5. Molecular constants from a GA calculation and an assigned fit of the rovibronic spectrum of the electronic origin of phenol−Ar.

Parameter	GA fit	Assigned fit
A'' (MHz)	1779.04(46)	1779.23(13)
B'' (MHz)	1119.40(24)	1119.315(76)
C'' (MHz)	904.82(9)	904.672(193)
T (K)	1.33	1.5
θ (°)	18.24(7)	20
ϕ (°)	25.76(93)	30
ΔA (MHz)	−43.43(28)	−43.516(42)
ΔB (MHz)	25.15(13)	25.191(31)
ΔC (MHz)	23.24(6)	23.227(20)

4.3.2. [7-D]Phenol−Ar

The phenol−Ar cluster is an example of a weakly van der Waals-bonded molecular cluster. Owing to the weak binding forces, centrifugal distortion (24, 25) might play a role in the determination of the molecular parameters. We included the five quartic centrifugal distortion constants for each electronic state in the fit. Compared to an equivalent fit without centrifugal constants, no significant improvement of the fit could be obtained. We performed the GA fit, where the azimuthal and polar angles of the transition dipole moment were allowed to vary between 0° and 90°, i.e., within their complete definition range. Further parameters to be varied were the rotational constants of ground and excited state, the center frequency,[4] and a single temperature. We used $\Delta w/\Delta_{lw} = 10$. As in the case of the [7-D]phenol−N_2 spectrum, the spectrum could be fitted without refining Δw. The resulting molecular parameters from the GA and from an assigned fit are presented in Table 5. While the rovibronic spectrum of phenol is of pure *b*-type, the argon atom, which is located above the aromatic ring and shifted slightly towards the hydroxy group, switches the axes, so that the transition moment in the cluster is oriented nearly along the inertial *c*-axis. From the parameters given in Table 5, we calculated the perpendicular r_0-distance of the argon atom to the aromatic ring with the program pKrFit (26). In the electronic ground state this distance is found to be 352.1 pm, while in the electronically excited state the distance is slightly reduced to 350.3 pm. Both values are in good agreement with distances found for other aromatic − noble gas clusters.

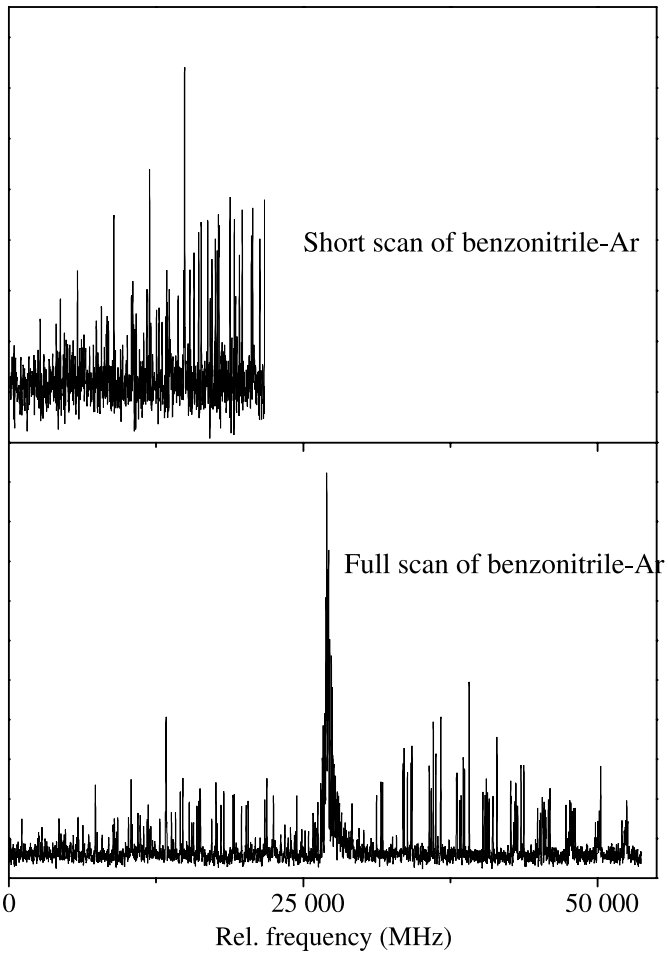
The GA fit of 10 molecular parameters was terminated in 6 min, using the sparse stick array implementation described in Sect. 3.3.2. Doubling of the dimension of the fitness surface by adding the 10 centrifugal distortion constants resulted in a computation time of 9 min. With the chosen ratio of $\Delta w/\Delta_{lw} = 10$, all five initial seeds converged to the same minimum within 500 generations.

4.3.3. Benzonitrile−Ar

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. As an example we chose the spectrum of the electronic origin of benzonitrile−Ar, shown in the upper trace of Fig. 4. Obviously, 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 6. The computing time was the same as for a complete spectrum discussed in the previous section. The electronic origin is found by the GA to be 8000 MHz to the blue of the high-frequency end of the spectrum. A GA fit to the complete spectrum with good signal-to-noise (lower trace in Fig. 4) yields slightly different molecular parameters (second column of Table 6). Nevertheless, the quality of the parameters obtained from the fit to the partial spectrum is surprisingly good. The only parameters that have relatively large deviations are the polar and azimuthal angles of the transition dipole moment. This is obviously due to the fact that the band type cannot be determined accurately from a fit of a single branch only.

Results from previous studies on benzonitrile−Ar are also given in the last column of Table 6. The current results for the excited state are more accurate than those of Helm et al. (1) because of the substantially lower resolution of their experiment.

Fig. 4. 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 the text.



[4]The value for the center frequency is omitted in all tables, as it is a relative number.

Table 6. Molecular constants from GA assignments of the partial spectrum of the origin of benzonitrile–Ar and the complete spectrum and from an assigned fit.

Parameter	GA fit[a]	GA fit[b]	Assigned fit	Other work
A'' (MHz)	1343.80(150)	1347.32(19)	1347.58(18)	1347.789(12)[c]
B'' (MHz)	1002.55(121)	1004.99(4)	1004.98(14)	1006.020(9)[c]
C'' (MHz)	717.68(108)	718.99(4)	717.70(39)	719.817(2)[c]
θ (°)	22(3)	17.53(7)	20	
ϕ (°)	82(5)	70.05(2)	70	
T(K)	1.71(3)	1.68(3)	2	
ΔA (MHz)	−32.89(44)	−32.61(3)	−32.47(23)	−33.8(36)[d]
ΔB (MHz)	20.60(30)	21.08(16)	20.87(14)	21.4(27)[d]
ΔC (MHz)	6.25(15)	6.76(7)	6.80(34)	10.2(75)[d]

Note: See text for details.

[a]Fit to the spectrum in the upper trace of Fig. 4.

[b]Fit to the spectrum in the lower trace of Fig. 4.

[c]From ref. 27.

[d]From ref. 1.

Our ground-state values do not completely agree with the very accurate microwave results from Dreizler and co-workers (27). This is an indication that the uncertainties in our parameters, based on the statistical behavior of the GA fits, are slightly underestimated.

4.4. Simultaneous GA fit of two 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. First of all, the number of transitions within a spectral interval is doubled, leading to very dense and congested spectra. Secondly, the number of molecular parameters is also doubled, which generates quite a large parameter space.

Overlapping spectra occur in particular if several isotopic species are investigated. While mass resolution techniques like resonance two-photon ionization with time-of-flight mass spectroscopy are able to separate the isotopic species with a different mass, the technique normally lacks experimental resolution owing to the pulse-width-limited resolution used in these studies. On the other hand, mass selection cannot be combined with high-resolution LIF spectra. In the next sections, we show that the GA spectrum assignments are capable of handling overlapping spectra both from different isotopomers as well as from different conformers.

4.4.1. [7-D][^{18}O]Phenol and [7-D][^{16}O]phenol

Further tests of the GA were performed with experimental spectra consisting of two sub-spectra, which originate from two isotopomers. As a first example, we chose the rovibronic spectrum of [^{18}O][7-D]phenol/[^{16}O][7-D]phenol, which had been assigned and published before (26). The isotopic enrichment of the phenol sample resulted in an isotopic purity of about 50% for the oxygen isotopes and of 100% for hydrogen. Since the spectral shift of the two spectra is about 3 GHz, the spectra completely overlap within the rovibronic contour. Both sub-spectra are of pure b-type, and thus the polar and azimuthal angles θ and ϕ do not need to be fit in this case. The rovibronic lines have a Voigt profile with a Gaussian line width of 20 MHz and a Lorentzian contribution of 12 MHz because of the fluorescence lifetime of 12.5 ns. The maximum J value in the calculation of the cost function is 15. Due to the smaller number of lines in the calculated spectrum, the computation time for the cost function drops drastically, and throughout all calculations on the phenol system, a population of 600 could be employed. The method for the GA evaluation employed is the same as described in Sect. 4.1. An initial fit was performed with a large value of $\Delta w/\Delta_{lw} = 15$. Table 7 gives the results of a GA fit, the limits of the molecular parameters used in the fit, and the results of a previously published assigned fit for comparison. Four GA evaluations with different starting values all converge to the same global minimum. Thus, $\Delta w/\Delta_{lw}$ was chosen correctly in the first step of the analysis. The comparison with the results from an assigned fit nevertheless shows deviations of about 2 MHz for the inertial parameters and their changes upon electronic excitation.

A second fit is performed with $\Delta w/\Delta_{lw} = 4$. The limits for the rotational constants and ΔA were reduced by a factor of 5 compared to the fit with $\Delta w/\Delta_{lw} = 15$, while the limits for ΔB and ΔC, which were already quite small, were reduced only by a factor of approximately 2. The results given in Table 8 clearly show not only that the GA fit converged to the global minimum, which is determined unambiguously from the assigned fit, but that the values of the inertial parameters are reproduced within their experimental accuracies. Figure 5 shows the experimental spectrum, along with the simulated spectra, using the parameters from the GA fit and from the assigned fit.

One run of the GA calculation with a population size of 600, a mutation probability of 0.05, and an elitism of 50% takes only about 10 min. This is due to the small J_{max} in the calculation of the simulated spectrum. With an initial seed of five different starting populations and two successive fits with different line width ratios and parameter spaces, a complete automated assignment was performed in less than 2 h.

In summary, in a two-step fit the GA evaluation succeeded in determining all molecular parameters for both completely overlapping spectra of [^{18}O][7-D]phenol and [^{16}O][7-D]phenol. The accuracy is comparable to an assigned fit of individual

Table 7. Fit of the inertial parameters of [^{18}O][7-D]phenol/[^{16}O][7-D]phenol using the GA along with the limits of the parameters used in the fit and the previously published parameters from an assigned fit of the individual line positions.

Parameter	GA fit	Lower limit	Upper limit	Assigned fit (26)
[^{18}O][7-D]Phenol				
A''	5 610.70	5 550.00	5 650.00	5 607.181(200)
B''	2 408.55	2 350.00	2 450.00	2 407.924(66)
C''	1 685.87	1 650.00	1 700.00	1 684.810(71)
ν_0	14 326.49	14 000.00	15 000.00	14 322.24(10)
ΔA	−334.60	−400.00	−300.00	−332.77(290)
ΔB	3.33	−5.00	5.00	4.02(120)
ΔC	−29.13	−50.00	0.00	−28.77(100)
[^{16}O][7-D]Phenol				
A''	5 604.84	5 550.00	5 650.00	5 608.222(90)
B''	2 528.93	2 500.00	2 550.00	2 527.965(57)
C''	1 743.30	1 700.00	1 800.00	1 742.737(40)
ν_0	17 032.26	16 000.00	18 000.00	17 031.00(10)
ΔA	−331.09	−400.00	−300.00	−332.56(160)
ΔB	3.18	−5.00	5.00	2.55(90)
ΔC	−31.87	−50.00	0.00	−31.51(70)
C_{fg}	4.496	—	—	2.580

Note: All values are given in MHz except C_{fg}, which is dimensionless. $\Delta w/\Delta_{lw} = 15$.

Table 8. Fit of the inertial parameters of [^{18}O][7-D]phenol/[^{16}O][7-D]phenol where the parameter space is narrowed compared to Table 7.

Parameter	GA fit	Lower limit	Upper limit	Assigned fit (26)
[^{18}O][7-D]Phenol				
A''	5 606.22	5 595.00	5 615.00	5 607.181(200)
B''	2 407.88	2 400.00	2 420.00	2 407.924(66)
C''	1 684.45	1 676.00	1 696.00	1 684.810(71)
ν_0	14 321.54	14 300.00	14 340.00	14 322.24(10)
ΔA	−332.02	−340.00	−320.00	−332.77(290)
ΔB	3.83	0.00	5.00	4.02(120)
ΔC	−28.53	−50.00	−20.00	−28.77(100)
[^{16}O][7-D]Phenol				
A''	5 608.53	5 595.00	5 615.00	5 608.222(90)
B''	2 528.02	2 520.00	2 540.00	2 527.965(57)
C''	1 742.62	1 735.00	1 755.00	1 742.737(40)
ν_0	17 033.63	17 000.00	17 050.00	17 031.00(10)
ΔA	−332.81	−340.00	−320.00	−332.56(160)
ΔB	2.54	0.00	5.00	2.55(90)
ΔC	−31.41	−50.00	−20.00	−31.51(70)
C_{fg}	2.342	—	—	2.580

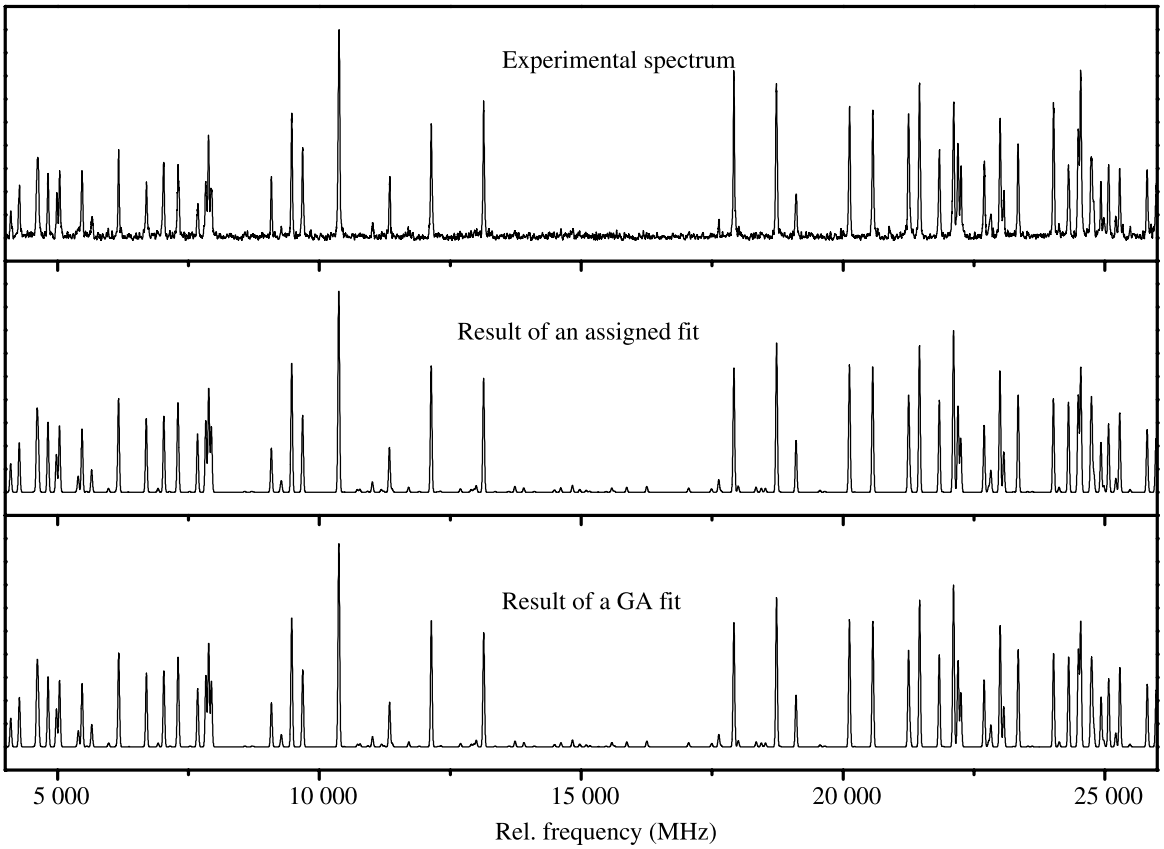
Note: All values are given in MHz except C_{fg}, which is dimensionless. $\Delta w/\Delta_{lw} = 4$.

rovibronic transitions. This evaluation is performed in less than 2 h, without any prior knowledge of the molecular parameters.

4.4.2. [3-D][7-D]Phenol and [5-D][7-D]phenol

Another example of overlapping electronic origins of different isotopomers of phenol is the pair [3-D][7-D]phenol and [5-D][7-D]phenol. Here, we have to assign simultaneously two bands with different Lorentzian widths in the Voigt profiles. In a recent publication (26), the line widths of both isotopomers were obtained from a fit of some individual rovibronic lines of each species (26). Since the GA performs a line shape fit of the complete spectrum, it should yield more accurate values. In a first step, the inertial parameters were determined using the GA with $\Delta w/\Delta_{lw} = 5$. For the determination of the line

Fig. 5. Experimental rovibronic spectrum of [^{18}O][7-D]phenol/[^{16}O][7-D]phenol, along with the simulation using the parameters from an assigned fit and from a GA fit. Intensities are given in arbitrary units.



shape parameters, the parameter space for the inertial parameters was narrowed down to 1 MHz and the GA fit was performed with $\Delta w = 0$. The Gaussian width was fixed to the experimentally determined value of 25 MHz. The temperature dependence of the relative intensity n is described by a two-temperature model (28):

$$[15] \qquad n(T_1, T_2, w_T) = \mathrm{e}^{-E/kT_1} + w_T \mathrm{e}^{-E/kT_2}$$

Here, E is the energy of the lower state, w_T a weight factor, and T_1 and T_2 the two temperatures. The intensity ratio between the spectra is fit as well. This resulted in improved values for the Lorentzian component of the line width. The resulting parameters are presented in Table 9 and compared with the values of an assigned fit. While the inertial parameters all agree within their uncertainties, the line width parameters are quite different. From the Lorentzian widths, the S_1-state lifetimes could be determined to be 23.5 ns for [3-D][7-D]phenol and 18.5 ns for [5-D][7-D]phenol. These values differ considerably from 38.8 ns and 15.6 ns obtained from a line shape fit to individual transitions in the spectrum. We attribute this difference to the limited number of single rovibronic transitions that could be used in the analysis of the line shapes, given in ref. 26.

As discussed before, the sparse stick spectrum array cannot be used for a fit of a line shape parameter. Instead, the sum in eq. [3] runs over all 80 700 data points, compared to just 745 lines with an intensity of more than 0.001 in the stick spectrum. This of course slows down the calculation of the fitness function considerably. A fit using the sparse array was finished in only 4.5 min, while the calculation with all data points needed 12 min.

4.4.3. Benzonitrile–^{20}Ne and benzonitrile–^{22}Ne

In all the cases discussed so far, the relative intensity of the two spectra is approximately 1:1. The situation is much more complicated if both isotopomers have very different abundances.

We performed a fit of the rovibronic spectrum of the isotopomeric pair benzonitrile–^{20}Ne/benzonitrile–^{22}Ne in the natural abundance of ^{20}Ne/^{22}Ne (91:9). In this case, the GA has the much more 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 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 (29), these monomer lines cannot be predicted with sufficient accuracy because they belong to very high J states.

The rovibronic spectrum of benzonitrile is of pure b-type. The neon atom is located above the aromatic ring, shifted slightly towards the cyano group. This structure gives rise to an axis switching. As a consequence, the transition moment in the cluster is oriented nearly along the inertial c-axis. For this molecular structure, an ac-hybrid is expected with strong c-type lines and much weaker a-type lines.

The results for the ac-hybrid type were checked against ab-, bc-, and abc-hybrid fits for consistency. The mean values for

Table 9. Molecular constants of [3-D][7-D]phenol and [5-D][7-D]phenol from the GA fit and an assigned fit (26) for comparison.

	[3-D][7-D]Phenol		[5-D][7-D]Phenol	
Parameter	GA	Assigned fit (26)	GA	Assigned fit (26)
A'' (MHz)	5337.98(79)	5338.161(166)	5349.82(8)	5349.789(175)
B'' (MHz)	2490.28(54)	2490.353(134)	2487.16(5)	2487.484(121)
C'' (MHz)	1698.58(31)	1698.567(65)	1697.98(4)	1698.088(76)
ΔA (MHz)	–309.32(8)	–309.42	–306.63(5)	–306.81
ΔB (MHz)	1.80(4)	1.69	0.31(5)	+0.08
ΔC (MHz)	–31.08(2)	–31.03	–31.26(3)	–31.13
T_1 (K)[a]	1.08	—	1.08	—
T_2 (K)[a]	2.81	—	2.81	—
w_T[a]	0.16	—	0.16	—
Ratio[b]	0.96	—	—	—
Δ_{Lorentz} (MHz)	8.57	10.2	6.77	4.1

[a]The assigned fit was performed with a one-temperature model.
[b]Ratio of intensities between the two spectra. No value given from the individual assigned fits.

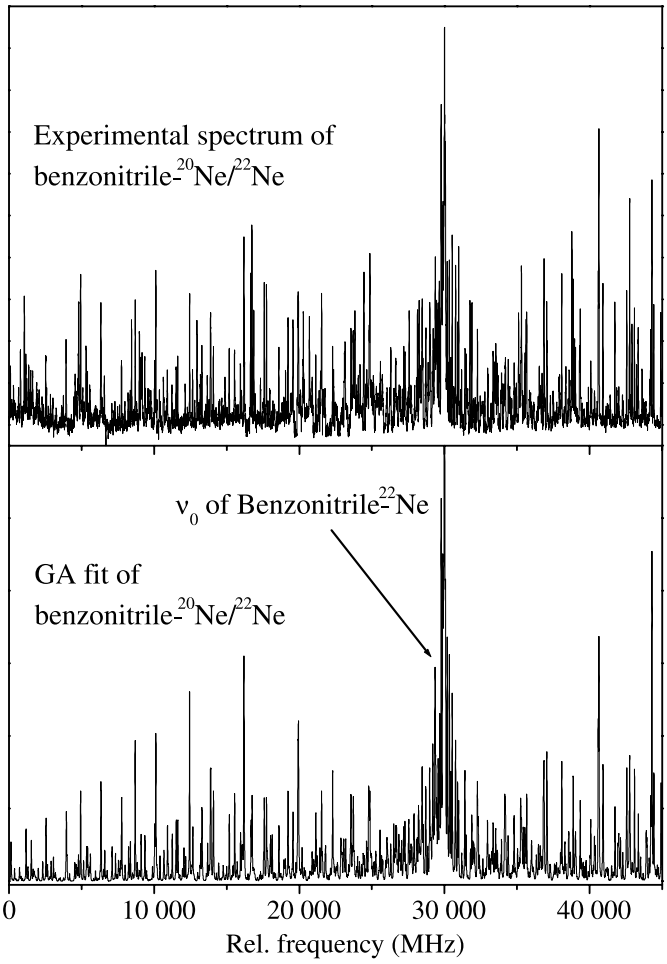
Table 10. Results of a GA assignment and values of the inertial parameters of benzonitrile–^{20}Ne/benzonitrile–^{22}Ne.

Parameter	Fit No. 1	Fit No. 2	Fit No. 3	Fit No. 4	Fit No. 5	Average
Benzonitrile–^{20}Ne						
A''	1 854.88	1 854.87	1 854.70	1 854.79	1 854.76	1 854.80
B''	1 193.99	1 193.80	1 194.06	1 194.11	1 194.02	1 193.99
C''	964.97	964.92	965.02	964.90	964.88	964.94
ν_0	30 327.45	30 328.14	30 325.94	30 325.97	30 328.11	30 327.12
ΔA	−2.83	−2.90	−2.78	−2.83	−2.88	−2.84
ΔB	−22.44	−22.41	−22.27	−22.32	−22.48	−22.39
ΔC	−22.44	−22.41	−22.27	−22.32	−22.48	−4.29
Benzonitrile–^{22}Ne						
A''	1 778.18	1 773.93	1 776.74	1 778.16	1 778.27	1 777.06
B''	1 178.15	1 178.14	1 165.85	1 178.86	1 177.87	1 175.77
C''	928.64	932.10	938.45	932.55	936.22	933.59
$\Delta\nu_0$	−1 454.95	−1 430.63	−1 450.33	−1 436.88	−1 461.91	−1 446.94
ΔA	25.75	26.28	26.13	24.93	25.40	25.70
ΔB	−16.05	−16.58	−13.43	−15.58	−13.59	−15.05
ΔC	−13.92	−14.35	−15.52	−14.43	−15.37	−14.72
C_{fg}	14.746	14.782	14.823	14.785	14.806	14.946

Note: All values are given in MHz except C_{fg}, which is dimensionles. $\Delta w/\Delta_{lw} = 1.5$. The five different fits were obtained with five different starting populations.

the cost functions are 7.95, 29.26, 9.27, and 7.18, respectively. Therefore, *ab*-hybrids can be discarded, which is in agreement with the geometry of the cluster. The *bc*-hybrid type fits slightly worse than the *ac*- type. However, the cost function differs only slightly because the spectrum is dominated by *c*-type lines. The *abc*-type did not improve the fit considerably. In conclusion, the initial assumption of the *ac*-hybrid type based on the approximate knowledge of the geometry was confirmed. Table 10 gives the molecular parameters obtained from a four-step GA fit. In the first step, $\Delta w/\Delta_{lw} = 10$ was employed. The search limits for the rotational constants were ±100 MHz for both isotopomers. The parameter limits were narrowed down to one-tenth that of the original size, centered around the best fit value of the first step. While the more abundant species (benzonitrile–^{20}Ne) 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–^{22}Ne is only one-tenth that 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. First, the parameter limits were reduced by only a factor of two, while $\Delta w/\Delta_{lw}$ was reduced to 7.5. In a subsequent step, $\Delta w/\Delta_{lw} = 5$ and limits of ±20 MHz for the rotational constants were employed. Finally, the molecular constants given in Table 10 were obtained for $\Delta w/\Delta_{lw} = 1.5$.

Fig. 6. Upper trace: experimental spectrum of benzonitrile–^{20}Ne/benzonitrile–^{22}Ne. Lower trace: simulation using the best parameters from Table 10. Intensities are given in arbitrary units.



In this case, the fit required some "fine tuning", which had to be done manually. Nevertheless, the results of the fit of the rovibronic spectrum of benzonitrile–^{20}Ne/benzonitrile–^{22}Ne (Fig. 6) 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.

5. Summary

In this paper we have shown that the GA is capable of treating a wide range of different spectra with complexity ranging from highly overlapping transitions to coinciding spectra of different isotopomers. Even 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.

If the spectrum under study originates from a single vibronic transition, convergence could be reached in a one-step fit with a typical value for $\Delta w/\Delta_{lw}$ of 10. A great enhancement is obtained in the accuracy of the line shape and intensity parameters. While the assigned fit always uses the information from a few selected single rovibronic lines, the GA utilizes all transitions in the spectrum to adapt these parameters. In particular, the Lorentzian width (lifetime), the temperature, and the orientation of the transition dipole moment in the molecule or in the molecular cluster are determined more accurately from GA fits.

In cases in which the spectrum is composed of two subspectra, a more advanced strategy has to be followed. A preliminary fit with a relatively broad weight function was performed and subsequently refined with both smaller weight function widths and narrowed parameter search space. Using this technique, very complicated spectra of two overlapping bands could be fitted using the GA. Even a large difference in intensity between the overlapping band forms no obstacle.

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. We even expect that in the case of small and (or) local perturbations, the main spectral features that conform to the model can be extracted and hence the perturbations are isolated.

The examples discussed demonstrate the extreme power of the GA in automated fitting and assigning of very complex spectra, spectra that can hardly be assigned and analyzed with conventional methods. It has been shown that the evaluation of the fitness function can be made to a minor contribution in the computing time if the particular properties of F_{fg} are fully exploited. 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.

Acknowledgements

We would like to thank Jos Hageman and Ron Wehrens for many helpful discussions. We thank Christian Ratzer for experimental help and Bert Groenenboom for giving the mathematical proof presented in Appendix A. The financial support of the Deutsche Forschungsgemeinschaft (SCHM 1043/9-4) is gratefully acknowledged. M.S. would like to thank the Nordrheinwestfälische Akademie der Wissenschaften for a grant, which made this work possible.

References

1. R.M. Helm, H.-P. Vogel, and H.J. Neusser. Chem. Phys. Lett. **270**, 285 (1997).
2. C.A. Haynam, D.V. Brumbaugh, and D.H. Levy. J. Chem. Phys. **81**, 2282 (1984).
3. L.A. Philips and D.H. Levy. J. Chem. Phys. **85**, 1327 (1986).
4. Y.R. Philips and D.H. Levy. J. Chem. Phys. **91**, 5278 (1989).
5. J.A. Hageman, R. Wehrens, R. de Gelder, W.L. Meerts, and L.M.C. Buydens. J. Chem. Phys. **113**, 7955 (2000).
6. R.E. Haupt and S.E. Haupt. Practical genetic algorithms. Wiley-Interscience, New York. 1988.
7. M. Mitchell. An introduction to genetic algorithms (complex adaptive systems). MIT Press, Cambridge, Mass. 1998.
8. M. Schmitt, J. Küpper, D. Spangenberg, and A. Westphal. Chem. Phys. **254**, 349 (2000).

9. M. Okruss, R. Müller, and A. Hese. J. Mol. Spectrosc. **193**, 293 (1999).
10. R.D. Suenram, F.J. Lovas, and G.T. Fraser. J. Mol. Spectrosc. **127**, 472 (1988).
11. W. Caminati and S. di Bernardo. J. Mol. Struct. **240**, 253 (1990).
12. S. Gerstenkorn and P. Luc. Atlas du spectre d'absoption de la molécule d'iode. CNRS, Paris. 1982.
13. H.C. Allen and P.C. Cross. Molecular vib-rotors. Wiley, New York. 1963.
14. S.C. Wang. Phys. Rev. **34**, 243 (1929).
15. D. Levine. PGAPack version 1.0 [computer program]. Available from ftp:ftp.mcs.anl.gov/pub/pgapck/pgapack.tar.z 1996.
16. J.H. Holland. Adaption in natural and artificial systems. MI: The University of Michigan Press, Ann-Arbor, Mich. 1975.
17. D.E. Goldberg. Genetic algorithms in search, optimisation and machine learning. Addison-Wesley, Reading, Mass. 1989.
18. I. Rechenberg. Evolutionsstrategie - Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Frommann-Holzboog, Stuttgart, Germany. 1973.
19. R. Wehrens, E. Pretsch, and L.M.C. Buydens. Anal. Chim. Acta, **388**, 265 (1999).
20. F. Gray. US Patent 2 632 058, 17 March 1953.
21. J. Schaffer, R. Caruana, L. Eshelman, and R. Das. *In* Proceedings of the third international conference on genetic algorithms. Morgan Kaufmann, San Mateo, Calif. 1989. p. 51.
22. M. Schmitt. C. Ratzer, and W.L. Meerts. J. Chem. Phys. **120**, 2752 (2004).
23. M.S. Ford, S.R. Haines, I. Pugliesi, C.E.H. Dessent, and K. Müller-Dethlefs. J. Electron Spectrosc. Relat. Phenom. **112**, 231 (2000).
24. J.K.G. Watson. J. Chem. Phys. **46**, 1935 (1967).
25. J.K.G. Watson. J. Chem. Phys. **48**, 4517 (1968).
26. C. Ratzer, J. Küpper, D. Spangenberg, and M. Schmitt. Chem. Phys. **283**, 153 (2002).
27. U. Dahmen, W. Stahl, and H. Dreizler. Ber. Bunsenges. Phys. Chem. **98**, 970 (1994).
28. Y.R. Wu and D.H. Levy. J. Chem. Phys. **91**, 5278 (1989).
29. D.R. Borst, T.M. Korter, and D.W. Pratt. Chem. Phys. Lett. **305**, 485 (2001).

Appendix A.

In this appendix we show that the real or complex matrix $\boldsymbol{W}$ with matrix elements $W_{ij} = w(r_i - r_j)$ is positive definite if $w(r)$ can be written as the inverse Fourier transform of a positive function $\tilde{w}(t)$. Let

$$[A.1] \quad w(r) = \int_{-\infty}^{\infty} \tilde{w}(t) e^{2\pi i r t} \mathrm{d}t$$

then

$$[A.2] \quad \boldsymbol{x}^{\dagger} \boldsymbol{W} \boldsymbol{x} = \sum_{i,j} x_i^* W_{ij} x_j$$

$$= \sum_{i,j} x_i^* \int_{-\infty}^{\infty} \tilde{w}(t) \exp\left[2\pi i (r_i - r_j) t\right] \mathrm{d}t \; x_j$$

$$= \int_{-\infty}^{\infty} \left| \sum_i x_i \exp\left[-2\pi i r_i t\right] \right|^2 \tilde{w}(t) \, \mathrm{d}t$$

Hence, if $\tilde{w}(t) > 0$ then $\boldsymbol{x}^{\dagger} \boldsymbol{W} \boldsymbol{x} > 0$ for any vector $\boldsymbol{x}$ and so $\boldsymbol{W}$ is positive definite. If $\tilde{w}(t)$ is zero for certain values of t, the integral is still positive. Note: one should read in eq. [4] $r_i = i$.