Fitting Fluorescence Spectra with Genetic Algorithms

J.A. Hageman[1], R. Wehrens[1], R. de Gelder[2], W. Leo Meerts[3], and L.M.C. Buydens[1*].

[1] Laboratory of Analytical Chemistry
[2] Department of Inorganic Chemistry
[3] Department of Molecular and Laser Physics
[*] Corresponding author

University of Nijmegen
Toernooiveld 1
6525 ED Nijmegen
The Netherlands

Abstract. Up until now it was not possible to automate the interpretation of spectra in which both frequencies and intensities contain (chemical) information. If the possibility of shifting peaks exist, pointwise comparisons of intensities at specific wavelengths is no longer adequate because wrong peaks could easily be compared. We show that with a suitable fitness function, which is generally applicable, spectra with shifted peaks can be solved using a standard GA. The method is very robust and illustrated using laser induced fluorescence spectra taken from Indole, Benzimidazole and 4-Aminobenzonitrile.

1 Introduction

Information on the identity, conformation or other physico chemical properties of a chemical sample is often obtained by one of several forms of spectroscopy. These techniques lead to a spectrum, which indicates the amount of energy absorbed or emitted (y-axis) and the frequencies at which this occurs (x-axis). Interpretation of these spectra then leads to the desired information. Several attempts to automate spectrum interpretation exist, but up to now they have only been successful in cases where the characteristic frequencies were known beforehand. Examples are originating from ^{1}H- and ^{13}C-NMR experiments [1, 2], IR-spectroscopy, UV-spectroscopy [3, 4], powder diffraction data [5–8] and fluorescence spectroscopy [9]. An often used procedure for the interpretation of spectra is the minimization of the differences between a theoretical and an experimental spectrum. The assumption is that parameter values, leading to a theoretical spectrum that is identical to the measured spectrum, are the correct values one is interested in. This procedure works correctly if only intensities of peaks need to be compared. As soon as peaks can change their position, these automatic approaches fail. This is the case in those flavors of spectroscopy where

the location of the peaks is related to the physical properties of the sample, e.g., nuclear magnetic resonance (NMR) and laser induced fluorescence (LIF). A pointwise comparison of peaks is in these cases no longer sensible, because the wrong peaks could easily be compared.

In the present paper, a method for the automated interpretation of such spectra is given. This method applies a Genetic Algorithm (GA) [10] with a fitness function able to deal with peak shifts. This is illustrated for the laser induced fluorescence (LIF) spectra of three organic molecules.

2 Laser Induced Fluorescence Spectroscopy

With LIF spectroscopy, it is possible to obtain information on the geometry of molecules in the form of rotational constants. These constants give information on intra- and intermolecular bond lengths and their changes upon excitation. Using a rigid asymmetric rotor Hamiltonian a theoretical spectrum can be calculated. The Hamiltonian describes the free rotation and accompanying energies and wave functions of a molecule in the gasfase. A full description of this Hamiltonian is beyond the scope of this paper but details can be found elsewhere [11]. The model is controlled by 12 parameters.

1. Six rotational constants. Three parameters (A", B", C") describing the ground state and three parameters (ΔA, ΔB, ΔC) describing the between the ground and excited state values (ΔA=(A'-A") etc. Here the double and single primes label the ground and excited state constants. These parameters are responsible for the location of (groups of) peaks and cause many peaks to shift left or right.
2. Three parameters (T_1, T_2 and W) that describe the relative intensities of the transitions between energy levels in a molecule described by the Hamiltonian.
3. Three further parameters: the line width ($\Delta\nu$), a frequency shift parameter (ν) and one parameter (θ) describing changes upon excitation of the molecule under investigation.

Minimization of the difference between a theoretical spectrum obtained with this model and an experimental spectrum should yield optimal values for the 12 parameters and in particular for the 6 rotational constants. In the next section, an adequate difference function is discussed. The power of this method is demonstrated for the LIF spectra of Indole, Benzimidazole and 4-Aminobenzonitril (4-ABN), which were discussed in Refs. [11, 12] and are shown in Figure 1.

3 Evaluation function

The similarity between the calculated spectrum and experimental spectrum has to be expressed in a single number if it is going to be used in combination with GAs. Several methods can be used for this purpose. In cases where only intensities of peaks can vary and frequencies (peak positions) remain constant, an

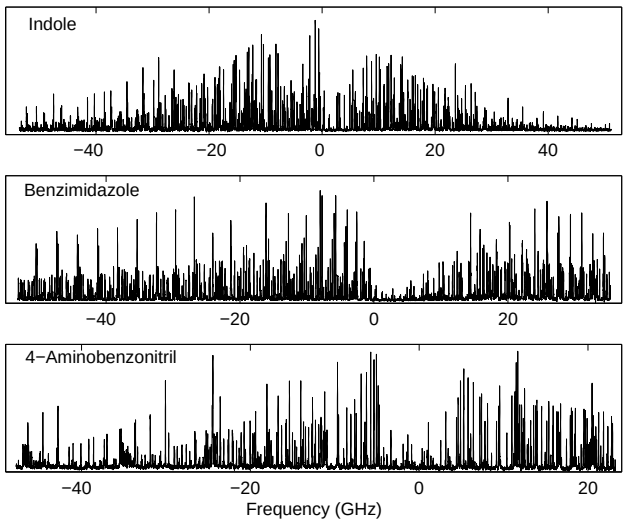

Fig. 1. High resolution LIF spectra of Indole, Benzimidazole and 4-ABN. In all cases the frequencies are relative to 0.0, according to Ref. [11] and Ref. [12], which corresponds to the rotational free electronic transition. The intensity on the vertical scale is in arbitrary units.

evaluation function based on a root-mean-square error (RMS) or a correlation will probably suffice for most applications [3,6–8]. If not only intensities but also frequencies contain information, comparison methods should include a comparison of the neighborhood to deal with peak shifts [5]. Another possibility would be to identify important and/or characteristic peaks frequencies and define a similarity measure with these peaks only.

Our initial attempts clearly demonstrated the inability of an RMS-type of evaluation function to recognize the similarity between spectra originating from nearly identical sets of parameters. Other approaches, based on peak picking and minimizing the distance to neighboring peaks in both spectra, failed as well. Since the relative position of peaks, in this application in particular, can change dramatically, one is never sure if the correct peak pairs are compared. With these types of evaluation functions, similar spectra with shifts in peak positions will not properly be recognized as similar.

To correctly compare this kind of spectra one should, in some way, compare the neighborhood of a given frequency. On way to achieve this is by using a cross correlation function as given in Eq. (1):

$$C_{fg}(r) = \frac{\sum_{x=0}^{x=k} f(x) \cdot g(x+r)}{\| f \| \cdot \| g \|} \tag{1}$$

Here *f(x)* and *g(x)* are spectra f and g with equal length k, and the term in the denominator is a normalization constant. Eq. (1) compares two spectra with a shift r added to one of them. When calculating several C_{fg} with different r values, it can be seen that similar spectra do not have their maximal value for $C_{fg}(0)$, as depicted in Figure 2. Here, the spectrum of Indole is compared with itself, two slightly modified spectra of Indole and a spectrum of another compound. The difference between the two different compounds is clearly visible, while the similarity between the three spectra is quite high.

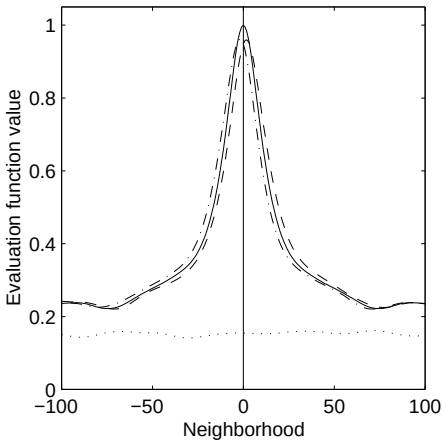

Fig. 2. Correlogram of the calculated spectrum of Indole. Autocorrelogram: solid line, Crosscorrelogram: dashed line (A" changed by 1.0 MHz), dash dotted line (ΔA changed by 1.0 MHz) and dotted line (calculated spectrum of Benzimidazole).

Larger shifts (r-values) can be penalized by a weight function *w(r)*. Several weight functions have been tested but the simple triangular function (Eq.(2)) worked best. Another advantage is that it is only controlled by one parameter, the base width of the triangle:

$$w(r) = 1 - \frac{\mid r \mid}{l} \tag{2}$$

Combining Eqs. (1) and (2) yields Eq. (3), the area under the weighted cross-correlation function:

$$C_{fg}^{ws} = \frac{\sum_{r=-l}^{r=l} C_{fg}(r)w(r)}{\sqrt{\sum_{r=-l}^{r=l} C_{ff}(r) * w(r)} * \sqrt{\sum_{r=-l}^{r=l} C_{gg}(r) * w(r)}} \tag{3}$$

For two identical spectra C_{fg}^{ws} is 1 and for distinctly different spectra C_{fg}^{ws} is close to zero.

A comparison of different evaluation function values is given in Table 1. It shows the inability of the RMS and, to a lesser extent, the correlation coefficient to recognize the similarity between the Indole spectra. All Indole spectra used in Table 1 are calculated with nearly identical sets of parameters and should therefor be recognized as similar which only is accomplished for our evaluation function. The correlation coefficient seems to work well, but is not good at discriminating between "very similar" and "similar" (data not shown).

Table 1. Evaluation values[a]

Evaluation function	A"	ΔA	Benzimidazole
RMS	1013.509	3389.386	95812.967
CORR	0.994	0.933	0.150
F	1.000	0.996	0.330

[a]Evaluation values calculated with different evaluation function (CORR = correlation coefficient, F = proposed function (left column). The calculated spectrum of Indole is compared with two nearly identical spectra of Indole (A" changed by 1.0 MHz, second column and ΔA changed by 1.0 MHz, third column), and the spectrum of Benzimidazole (right column).

The final evaluation function used in the GA calculations is defined as:

$$F = 100 * (1 - C_{fg}^{ws}) \tag{4}$$

and its value is minimized.

A more detailed discussion and comparisons with other methods for the assessment of similarity between 1-dimensional spectra, can be found in the work of De Gelder et al. [13].

4 Experimental

The spectra of Indole, Benzimidazole and 4-ABN are shown in Figure 1. The spectra of Indole and Benzimidazole contain 65536 equidistant data points and the spectrum of 4-ABN contains 40972 data points. All 12 parameters were coded as 10-bit gray binary numbers. The rotational constants in the excited state are expressed on the string as the difference with the ground state. T_2 is coded on the string as α, with $T_2 = \alpha * T_1$ and $\alpha > 1$. The calculated spectra always contain the same number of data points as the corresponding experimental ones. The optimal settings of the GA were determined in preliminary experiments, based on previous experience, and are shown in Table 2.

The optimal size of the neighborhood in Eq. (2) has been established from several experiments. The optimal value for l was 100 data points. A larger range

Table 2. GA settings.

Setting	Value
maximum number of generations	500
population size	300
elitism	150
crossover type	two-point-crossover
crossover probability	0.85
mutation type	new random value within boundaries
mutation probability	0.05
selection type	probabilistic
fitness type	raw[a]

[a]Fitness value increases inversely proportional with evaluation value of a string.

also results in a correct solution but leads to longer run times. For a significantly smaller range no correct solution is obtained, indicating that the inclusion of neighborhood information is crucial. After establishing the optimal settings, the experimental spectra of Indole, Benzimidazole and 4-ABN were fitted using boundary constraints as given in Table 3. The duration of a run has been set to 500 generations, long enough to converge to a minimum. All runs were repeated 5 times with different random generator seeds.

The robustness of the GA method was investigated in a number of runs. Synthetic spectra of Indole and Benzimidazole were modified with different levels of normally distributed (white) noise, increased line widths and a combination of these two factors.

All GA calculations were performed with the GA library PGAPack version 1.0 [14], which can run on parallel processors. PGAPack and the evaluation function are written in ANSI-C, the rigid asymmetric rotor Hamiltonian function was written in Fortran. All calculations were performed on a Sun-Ultra-Enterprise-10000 with 24 processors each running at 333 MHz. With 16 processors, the average runtime was about half an hour for 500 generations and 65536 data points. In practice, this runtime can be reduced drastically, because often runs converged to their final solution long before the maximum number of generations was reached.

5 Results and Discussion

Table 4 shows the 12 parameters for all four experimental spectra resulting from the GA, together with the results reported by Ref. [11] (Indole and Benzimidazole) and Ref. [12] (4-ABN), using the manual methods. The values obtained with our present GA approach are in close agreement with these previous results. Results from a GA using an evaluation function based on the RMS did not lead to valid results at all. The correlation function leads to improved results, but still was not able to fit all 12 parameters. A comparison of the error landscapes

Table 3. Boundary constraints for all 12 parameters used for Indole, Benzimidazole and 4-ABN[a].

Parameter	Boundary constraints	
	Indole and Benzimidazole	4-ABN
A"	3800 - 4200	5000 - 6000
B"	1400 - 1800	800 - 1200
C"	800 - 1400	600 - 1000
T_1	1 - 6[b]	1 - 6
T_2^d	1.5 - 5	1.5 - 5
W	0 - 1	0 - 1
θ	0° - 90°	90°, fixed[e]
ν	-300 - 300[c]	-5000 - 5000
ΔA	-200 - 0	-400 - 400
ΔB	-50 - 0	-100 - 100
ΔC	-50 - 0	-100 - 100
$\Delta\nu$	10 - 40	10 - 90

[a]Rotational constants in the ground state are indicated by A", B" and C". Rotational constants in the excited state are given by their deviations from the ground state (ΔA, ΔB and ΔC). $\Delta\nu$ is line width of the Lorentzian peaks. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1 and T_2 in K.
[b]Range is 2 - 8 for the spectrum taken from Benzimidazole.
[c]The frequency of the origin (ν) is set to zero. The area of deviation is taken to be $\pm$ 10% of the reported value from Refs. [11] [12]
[d]$T_2=\alpha*T_1$ where α has been optimized with the constrained $\alpha > 1$
[e]Determined by the geometry of the molecule.

of the RMS, correlation and weighted crosscorrelation is depicted in Figure 3. In all three plots, ΔA and ΔB are varied over a grid covering the complete range, while the remaining parameters are held fixed at their optimal values. The effect of the correlation function in comparison with the RMS based function is a smoothing of the error landscape, which reduces the number of local minima and should make it easier for the GA to locate the global optimum. The smoothing effect for our new evaluation function (F) is much larger compared to that of the correlation function. Moreover, the width of the "basis of attraction" is enlarged significantly.

The GA using the proposed evaluation function was able to find the correct solution for the Indole and Benzimidazole spectra in all 5 replicated runs. The correct solution for the 4-ABN data was found in only 2 of the 5 cases, as shown in Figure 4. The cause of the reduced reproducibility of the 4-ABN run is probably the larger boundary constraints, which makes it more difficult for the GA to locate the correct solution.

The absolute evaluation function values did not reach the same level for the 3 compounds. This is due to the noise level, line width and total number of data points in a particular spectrum. High noise levels intrinsically give rise to large

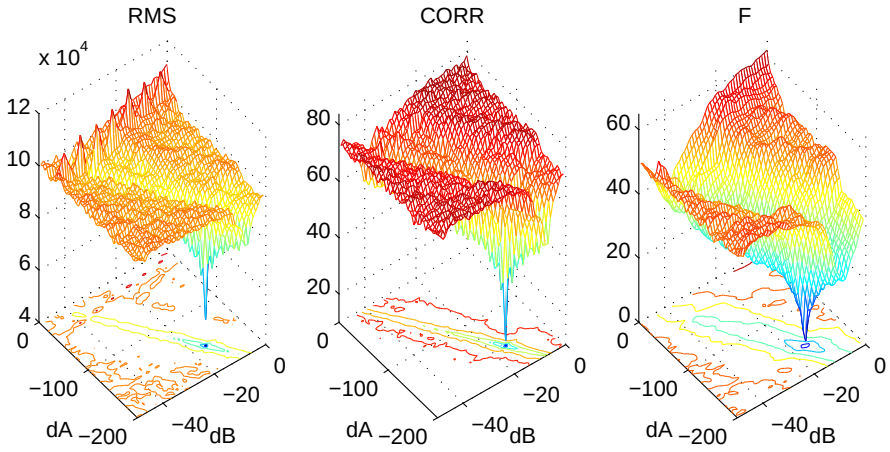

Fig. 3. Difference in error landscape between a RMS-based (left), correlation coefficient (middle) and our new evaluation function (right).

values of the evaluation function. However, the minimum obtained in each case is the global minimum, irrespective of the absolute evaluation value.

The addition of noise and increased line widths to the spectra led to an increase in evaluation value. Although the quality of the solutions appeared to deteriorate, the rotational constants were hardly influenced by the elevated noise levels. The deviations were mostly found in T_1, T_2 and in θ. Because one is mostly interested in the rotational constants the method can be considered quite robust for the determination of these parameters.

A decrease of the number of data points (where the overall frequency range is kept constant) only shows an effect on the Benzimidazole spectrum. For a smaller number of data points, the solutions become worse. This is due to the

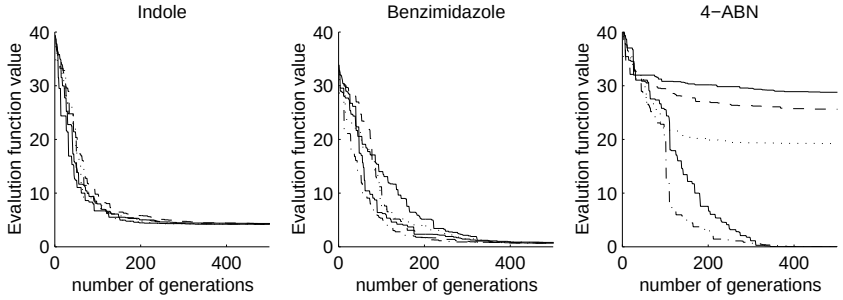

Fig. 4. Progression of the best solution during a run for Indole, Benzimidazole and 4-ABN.

Table 4. Results from GA runs for Indole, Benzimidazole and 4-ABN.[a]

	Indole		Benzimidazole		4-ABN	
	GA	Ref [11]	GA	Ref [11]	GA	Ref [12]
A"	3879.8	3880.7	3929.0	3930.5	5579.7	5579.3
B"	1637.0	1637.5	1679.2	1679.5	990.23	990.26
C"	1151.3	1152.1	1177.1	1176.7	841.45	841.39
T_1	2.2	1.50	5.63	4.88	2.63	3
T_2	7.93	5.03	21.52	20.0	4.56	-
W	0.1	0.22	0.423	0.42	0.84	-
θ	37.4°	±38.3	22.0°	±22.0°	0°	0°
ν^b	0.78	0.0	1.04	0.0	-1.61	0.0
ΔA	-134.70	134.66	-155.62	-155.70	-315.54	-316.61
ΔB	-18.08	-17.96	-15.30	-15.37	10.66	10.849
ΔC	-20.72	-20.77	-21.41	-21.31	0.29	0.095
$\Delta\nu$	16.2	20.05	19.33	19.45	16.16	26
			Evaluation Values			
best	4.1815	9.956	0.6460	0.56	1.200	2.800

[a] All runs were repeated 5 times with different seeds for the random number generator; the solutions with the lowest evaluation values are shown. Values from Ref. [11] and Ref. [12] are listed in the respective columns. Molecular constants in Ref. [11] are averages from multiple spectra and were determined using very accurate ground rotational constants from microwave experiments. Values given here are based on a spectral analysis of the same spectrum, where the ground rotational constants were also determined. The parameters that describe the relative intensity of a transition (T_1, T_2, W) have different values from those reported in Ref. [11]. (Ref. [12] used a one-temperature model, so their findings cannot be compared with our results). The difference is due to the fact that different sets of parameters result in equal spectral intensities. Rotational constants, ν and $\Delta\nu$ are in MHz, T_1 and T_2 in K.
[b] The absolute frequency of the origin is given as the deviation from the reported value from Ref. [11] and Ref. [12].

fact that spectral information gets lost if the distance between two successive data points becomes too large.

6 Conclusions

The automated interpretation of high resolution spectra becomes of great importance if the interpretation by other methods is not feasible, is too time-consuming or just tedious. In cases where intensities as well as frequencies are dependant on the parameters that are optimized, it is crucial that both are taken into account when devising an appropriate difference function. In our example, the success of the GA method crucially depends on the newly developed evaluation function. Other, more standard, evaluation functions lead to no results.

The GA method is quite robust. It is insensitive to large line widths in the spectrum, and only at very high noise levels do the results deteriorate. Even

then, the most important parameters were unaffected. It is shown that the GA is able to use all information present in the spectrum and therefore its performance increases with the number of data points. The method of matching experimental data with simulated model data taking into account peak opens up vast possibilities in other fields, such as NMR-spectroscopy, where peak shifts determine spectral characteristics.

References

1. D. S. Stephenson and Gerhard Binsch. Automated analysis of high-resolution NMR spectra. I. Principles and computational strategy. *J. Magn. Reson.*, 37:395–407, 1980.
2. W.Y. Choy and B.C. Sanctuary. Using genetic algorithms with a priori knowledge for quantitative NMR signal analysis. *J. Chem. Inf. Comput. Sci.*, 38:685–690, 1998.
3. J. Dods, D. Gruner, and P. Brumer. A genetic algorithm approach to fitting polyatomic spectra via geometry shifts. *Chem. Phys. Lett.*, 261:612–619, 1996.
4. R.M. Helm, H.-P. Vogel, and H.J. Neusser. Highly resolved UV spectroscopy: structure of S_1 benzonitrile and benzonitrile-argon by correlation automated rotational fitting. *Chem. Phys. Lett.*, 270:285–292, 1977.
5. H.R. Karfunkel, B. Rohde, F.J.J. Leusen, R.J. Gdanitz, and G. Rihs. Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comp. Chem.*, 14(10):1125–1135, 1993.
6. K. Shankland, W.I.F. David, and T. Csoka. Crystal structure determination from powder diffraction data by the application of a genetic algorithm. *Z. Kristall.*, 212:550–552, 1997.
7. K.D.M. Harris, R.L. Johnston, and B.M. Kariuki. The genetic algorithm: Foundations and applications in structure solution from powder diffraction data. *Acta Cryst.*, A54:632–645, 1998.
8. B.M. Kariuki, S.A. Belmonte, M. I. McMahon, R.L. Johnston, D.M. Harris, and R.J. Nelmes. A new approach for indezing powder diffraction data based on whole-profile fitting and global optimization using a genetic algorithm. *J. Synchrotron Rad.*, 6:87–92, 1999.
9. M. P. Jacobson, S.L. Coy, and R.W. Field. Extended cross correlation: A technique for spectroscopic pattern recognition. *J. Chem. Phys.*, 107(20):8349–8356, 1997.
10. R. Wehrens and L.M.C. Buydens. Evolutionary optimisation: a tutorial. *Trends in Analytical Chemistry*, 17(4):193–203, 1998.
11. G. Berden, W.L. Meerts, and E. Jalviste. Rotationally resolved ultraviolet spectroscopy of indole, indazole and benzimidazole: Inertial axis reorientation in the $S_1(^1L_b) \leftarrow S_0$ transitions. *J. Chem. Phys.*, 103(22):9596–9606, 1995.
12. G. Berden, J. Van Rooy, W.L. Meerts, and Z.A. Zachariasse. *Chem. Phys. Lett.*, 278:373, 1997.
13. R. de. Gelder, R. Wehrens, J.A. Hageman, and L.M.C. Buydens. A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. submitted.
14. D. Levine. PGAPack V1.0. PGAPack can be obtained from anonymous ftp from: ftp://ftp.mcs.anl.gov/pub/pgapack/pgapack.tar.Z.