A new genetic algorithm to be used in the direct fit of potential energy curves to *ab initio* and spectroscopic data

2008 J. Phys. B: At. Mol. Opt. Phys. 41 085103

(http://iopscience.iop.org/0953-4075/41/8/085103)

J. Phys. B: At. Mol. Opt. Phys. **41** (2008) 085103 (15pp) doi:10.1088/0953-4075/41/8/085103

A new genetic algorithm to be used in the direct fit of potential energy curves to *ab initio* and spectroscopic data

J M C Marques[1], F V Prudente[2], F B Pereira[3,4], M M Almeida[2], A M Maniero[5] and C E Fellows[6]

[1] Departamento de Química, Universidade de Coimbra, 3004-535 Coimbra, Portugal
[2] Instituto de Física, Universidade Federal da Bahia, 40210-340 Salvador, BA, Brazil
[3] Instituto Superior de Engenharia de Coimbra, Quinta da Nora, 3030-199 Coimbra, Portugal
[4] Centro de Informática e Sistemas da Universidade de Coimbra (CISUC), 3030-290 Coimbra, Portugal
[5] Instituto de Ciências Ambientais e Desenvolvimento Sustentável, Universidade Federal da Bahia, 47805-100 Barreiras, BA, Brazil
[6] Instituto de Física, Universidade Federal Fluminense, 24210-346 Niterói, RJ, Brazil

E-mail: qtmarque@ci.uc.pt

Received 30 October 2007, in final form 13 February 2008
Published 7 April 2008
Online at stacks.iop.org/JPhysB/41/085103

Abstract

We propose a two-step genetic algorithm (GA) to fit potential energy curves to both *ab initio* and spectroscopic data. In the first step, the GA is applied to fit only the *ab initio* points; the parameters of the potential so obtained are then used in the second-step GA optimization, where both *ab initio* and spectroscopic data are included in the fitting procedure. We have tested this methodology for the extended-Rydberg function, but it can be applied to other functions providing they are sufficiently flexible to fit the data. The results for NaLi and Ar_2 diatomic molecules show that the present method provides an efficient way to obtain diatomic potentials with spectroscopic accuracy.

1. Introduction

For a long time, spectroscopic data of diatomic molecules have been used to obtain Rydberg–Klein–Rees (RKR) potential points through a direct semi-classical inversion method [1–3]. These potential points can then be used in a least-squares fit to an appropriate functional form to obtain the interaction energy as a function of the interatomic distance. Another source of data that may be used altogether with the RKR points to build in potential energy curves arises from the inverse perturbation approach (IPA) proposed by Kosman and Hinze [4] and by Vidal and Scheingraber [5]. In opposition to the semi-classical RKR method, IPA has been established as a fully quantum-mechanical approach to obtain potential energy curves of diatomics. However, both RKR and IPA methods are based on an inversion process of the spectrum data to obtain sets of points (i.e., energy versus interatomic distance) belonging to the potential curve, and hence cannot be seen, rigorously, as direct spectroscopic information. Nonetheless, the joint use of these kinds of data and rovibrational energies has been recently carried out [6] in a direct least-squares fit of diatomic potentials by employing a standard Levenberg–Marquardt algorithm as implemented in the MINPACK package [7].

In the last decade or so, there has been a great interest in the application of direct least-squares fitting procedures to obtain diatomic potential energy curves from large sets of spectroscopic data. Basically, such methods consist of fitting the parameters of the radial Hamiltonians associated with the electronic states of interest so that the experimental (vibrational–rotational and pure rotational) spectrum line positions are accurately reproduced by the corresponding eigenvalues. In addition, the associated atomic-mass-dependent Born–Oppenheimer breakdown (BOB) radial strength functions may be also obtained through this direct fitting procedure providing that data are available for various isotopologues. It is worth noting in this context the pioneering work of Coxon and Hajigeorgiou [8–10] and Zimmermann and collaborators [11] that first developed numerical methodologies to perform the direct fits. Since then, many other groups [12–22] have explored and expanded

upon the advantages of such methods. In turn, an algebraic approach has been suggested by Ogilvie [23, 24] to deal with the reduction of diatomic spectrum data to a compact set of parameters defining both potential energy and BOB radial functions. Though over the years the numerical and the algebraic approaches have led to some apparent discrepancies between each other, Le Roy [17] has recently demonstrated that the source of such disagreement is simply due to the truncation convention adopted by Ogilvie for the Dunham expansion as it is implemented in the RADIATOM program [23–30].

Although electronic structure calculations are nowadays affordable for many diatomic systems, most of the available *ab initio* potentials cannot still reproduce the experimental vibrational levels with spectroscopic accuracy. In particular, some of the present authors [31] have employed multi-configurational quasi-degenerate perturbation theory [32, 33] (MCQDPT), which uses the MCSCF orbitals as reference functions in the perturbation procedure (as implemented in GAMMESS-US package [34, 35]), to calculate the NaLi potential curves for some singlet ($X^1\Sigma^+$ and $A^1\Sigma^+$) and triplet ($a^3\Sigma^+$, $b^3\Pi^+$ and $c^3\Sigma^+$) electronic states, as well as the transition dipole momenta among states of the same symmetry; the *ab initio* points were then interpolated to calculate the vibrational levels and the radiative transition probabilities by using the LEVEL program [36]. For example, the vibrational levels for NaLi($X^1\Sigma^+$) present discrepancies with the experimental values [37, 38] that vary from less than 1 cm^{-1} for $v = 0$ up to $\sim$18 cm^{-1} for $v = 24$.

Another kind of system that represents a challenge for theoreticians is the rare-gas dimers. The diatomic potential of such systems is characterized by a shallow and small van der Waals minimum, and is dominated by dispersion interactions at large interatomic distances. Because of this, highly correlated *ab initio* methods and basis sets with a large number of polarization and diffuse functions must be used to get an accurate potential energy curve for rare-gas dimers. In the case of Ar_2 (which is a benchmark system in many contexts) the problem is even more acute due to the large number of electrons that should be accounted for in the electronic structure calculation. However, the enormous progress carried out on this field (e.g., by the use couple-cluster CCSD(T) methods ([39] and references therein), augmented correlation-consistent basis sets [40, 41], bond functions [42–44] and extrapolating results to the complete basis-set (CBS) limit [45, 46]) led to the calculation of probably the most accurate Ar_2 potential [47, 48], which can reproduce, among other properties, the five experimental $v' \rightarrow v''$ transitions [49] within less than 1 cm^{-1}.

Since the beginning of the last decade, genetic algorithms (GAs) have been applied to solve several optimization problems in chemistry and physics. Indeed, GAs have shown themselves to be a powerful tool for searching the global optimum in difficult cases such as atomic clusters geometry optimization (see [50] and references therein) and protein folding [51, 52]. Although most of the work on this topic has been devoted to build in efficient GAs for discovering the global minimum structure of molecules and clusters [53–60], there have been also applications to deal with the fitting process of experimental data (see [61, 62] and references therein). Due to its difficulty, there are advantages in applying GAs in the tedious task of spectral assignment (which was traditionally done by visual identification of patterns) and, thus, automate the process in an efficient way [63–66]. Specifically, Hageman *et al* [63] have shown that GAs can be used to determine rotational constants from rovibronic spectra. Such a methodology makes use of the rigid asymmetric rotor Hamiltonian model to obtain a theoretical rovibronic spectrum, while the differences between this and the experimental spectrum are, then, minimized by applying the GA. Meerts and Schmitt [66] have demonstrated the efficiency of this method on solving the assignment problem in various complex cases, e.g., spectra with high overlapping transitions, coinciding spectra of different isotopomers and life-time broadened spectra. In addition, GAs have been successfully applied to fit other spectroscopic data: nuclear magnetic resonance [67], fluorescence/absorption spectra in polyatomic molecules [68], Mössbauer spectroscopy [69], multi-objective x-ray spectroscopic analysis [70] and powder electron paramagnetic resonance (EPR) spectra [71]. In particular, Hennessy and Kelley [61] have applied real-valued multi-objective GAs to solve the spectral inversion problem for sets of absorption spectra and resonance Raman excitation profiles. In turn, Spalek *et al* [71] have proposed the joint use of a GA with the Powell method to perform a least-squares fit of the powder EPR spectrum data.

The application of GAs to the difficult task of discovering the set of nonlinear parameters of a potential energy function that best fits theoretical and/or experimental data have motivated also the interest of researchers [72–75]. Recently, a GA using a binary representation has been applied to fit the potential energy surface of NaHF [74] and the potential curves of both H_2^+ and Li_2 [75]. The extended-Rydberg function has been used as a model to fit these diatomic systems to accurate *ab initio* data, but they cannot reproduce the corresponding spectra with an accuracy below 1 cm^{-1}; in fact, discrepancies as large as 5.79 cm^{-1} and 31.45 cm^{-1} are observed [75] for H_2^+ and Li_2, respectively.

In this work, we want to investigate the possibility of using a GA to fit diatomic potential curves to both *ab initio* energies and experimental vibrational levels. For this purpose, we suggest and analyse a hybrid real-valued GA which fits the *ab initio* data to get a first estimate of the parameters which will then be used in the joint fit of the *ab initio* and vibrational levels. Since the major goal of the present work stems from testing the ability of the new GA to fit accurately diatomic spectroscopic data, no attempt has been made to use state-of-the-art *ab initio* points (e.g., calculated by including relativistic effects and BOB corrections). In fact, by including *ab initio* data, we just want to guarantee the preservation of the form of the potential curve during the fitting procedure in the absence of a large set of experimental data. However, the effect of fitting only the vibrational levels (i.e., without including *ab initio* data) was also analysed here for NaLi. Moreover, we are interested on the possibility of a simple model function reproducing the diatomic spectra (i.e., data reduction) and,

hence, we have no concern about extrapolating realistically at large and small internuclear distances. Thus, we apply the extended-Rydberg function to model the Ar_2 and NaLi potentials and analyse the improving in the fitting against increasing the number of parameters included in the model. The plan of this paper is as follows. In section 2, we describe the genetic algorithm (GA) and the fitting procedure proposed in this work. A discussion about the performance of the GA, as well as the results for fitting the spectra of Ar_2 and NaLi are given in section 3. Conclusions and final remarks are in section 4.

2. Methodology

We aim to obtain analytic potential energy curves with spectroscopy accuracy by performing a nonlinear least-squares fit to both *ab initio* and vibrational spectrum information. In other words, this can be seen as the general problem to find the best set of parameters $\mathbf{z} \equiv (z_1, z_2, \ldots, z_n)$ which minimizes an objective function

$$\chi^2(\mathbf{z}) = \sum_{j=1}^{N} \omega_j [f(x_j; \mathbf{z}) - y_j]^2, \qquad (1)$$

where the pair (x_j, y_j) designates the point j of the set of N data points of the fit, ω_j is the weight given to point j and $f(x; \mathbf{z})$ is a model function employed in the fitting procedure.

In this work, we propose to employ a hybrid GA, where the evolutionary algorithm is complemented by a local search procedure. Its task is to discover the global minimum of equation (1). Basically, we want to find out the values that provide the best least-squares fit of a model function $f(x_j; \mathbf{z})$ to a data set of N points y_j. According to the standard GA terminology, each parameter z_i is a gene, a specific set of parameters $\mathbf{z}$ (i.e., a solution) is an individual and a group of individuals forms the population. The iterative processing of successive populations is guided by the two fundamental forces of evolutionary systems: a probabilistic selection mechanism gives preference to high-quality individuals from the current population and, afterwards, variation operators are applied to selected solutions to obtain the descendants. The goal of the evolution is to determine the best nonlinear least-squares fit.

In the following subsection, we present the details of the hybrid GA and the direct fitting procedure is discussed in subsection 2.2.

2.1. Genetic algorithm

The flow chart in figure 1 presents an overview of the main steps performed by the hybrid GA. In the beginning of the optimization, the initial population is generated by randomly selecting a set of solutions from the search space. The composition of the search space is established by selecting an appropriate interval for each one of the fitting parameters or genes. Figure 2(a) graphically represents the structure of a population: a set of K solutions (K represents the population size), where each individual encodes a specific set of parameters $(z_1, z_2, \ldots, z_n)$. The Broyden–Fletcher–Goldfarb–Shanno limited memory quasi-Newton (L-BFGS) method [76, 77] is then applied. The application of this method aims to

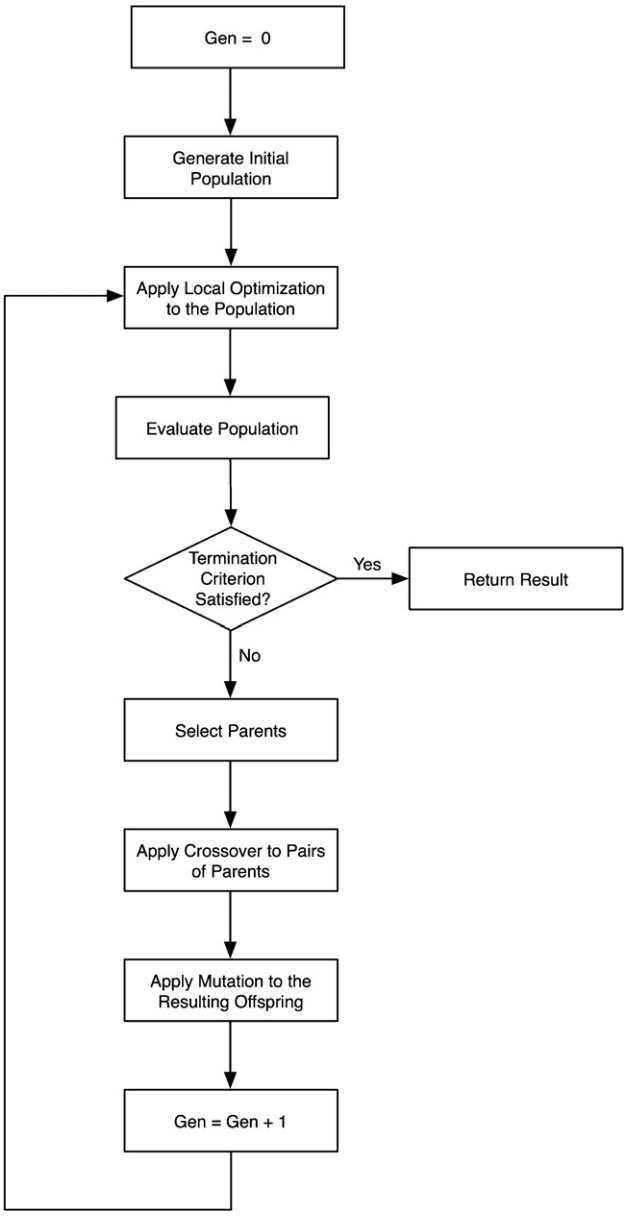

Figure 1. Flowchart of the genetic algorithm used in the present work.

perform a local improvement on each one of the solutions that belong to the current population. During local optimization, the maximum number of iterations that can be performed is specified by a parameter of the algorithm, the local search length (LSL). However, L-BFGS stops as soon as it finds a local optimum, so the effective number of iterations can be smaller than the value specified by LSL. Note that L-BFGS is a powerful optimization technique which combines the modest storage and computational requirements of conjugate gradient methods with the superlinear convergence exhibited by full memory quasi-Newton methods; it only requires the function (i.e., χ^2) and its gradient.

After local optimization, the fitness of each individual is evaluated with equation (1). The quality of a solution reflects how well its set of fitting parameters enable the adjustment of

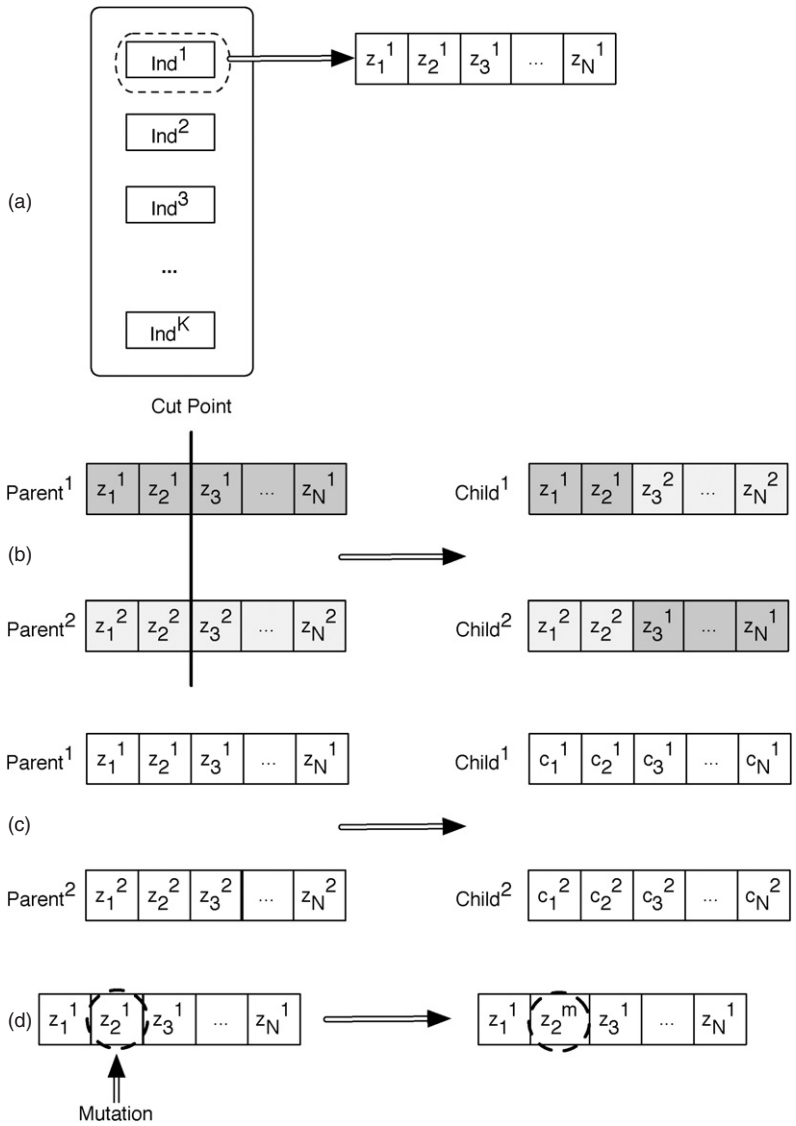

Figure 2. Population and main operators used in the present GA: (a) structure of the population; (b) one-point crossover; (c) SBX crossover; (d) mutation (see the text).

the data points to the model function ($f(x_j; \mathbf{z})$). Since this is a minimization problem, individuals with low fitness encode better solutions.

As this GA adopts a fully generational model, in each generation the whole population is replaced by its offspring. The new generation (i.e., the new set of solutions) is obtained after a sequence of steps. The first step is to select a set of parents from the individuals that compose the current generation. Tournament selection is the method adopted in this work. To select one parent, this mechanism performs the following operations: it randomly chooses T individuals from the current population (T is a parameter of the algorithm called the tourney size). The best solution from this set is selected as a parent of the new generation. For a population with K individuals, the procedure is repeated K times.

The standard genetic operators, crossover and mutation, are then applied to the selected individuals to obtain the descendants. First, crossover is applied, with a given probability, to all pairs of parents. Afterwards, mutation is applied, with a given probability, to each gene of the resulting individuals.

In what concerns crossover, two different operators are considered: one-point crossover and simulated binary crossover (SBX) [78]. The first one is an example of a discrete recombination operator. It selects a random cut point along the sequence of fitting parameters and the two parents involved in the operation are detached after this location. The two children are obtained by exchanging the tails. Figure 2(b) illustrates how this operator acts. In this example the cut point is applied after the second gene. The main disadvantage of discrete crossover operators is that they do not insert new values into the population (they just rearrange existing values). To overcome this limitation there are alternative operators known as arithmetic crossovers. When they are used, the value for each fitting parameter from the children is calculated as a

weighted average of the corresponding values from the parents. SBX is one well-known operator belonging to this group and it has achieved good results in a number of real-valued problems of varying difficulty and dimensionality [78, 79]. Figure 2(c) illustrates how this operator acts. The values for the parameters that will compose the children solutions $\{\text{Child}^1, \text{Child}^2\}$ are obtained from the parent solutions $\{\text{Parent}^1, \text{Parent}^2\}$ in the the following way [79]:

(1) select a random value $\mu \in [0, 1]$,
(2) calculate

$$\beta = (2\mu)^{1/(\eta+1)}, \quad \text{if} \quad \mu \leqslant 0.5$$
$$\beta = (1/2(1-\mu))^{1/(\eta+1)}, \quad \text{if} \quad \mu > 0.5,$$

(3) Obtain children

$$\text{Child}^1 = 0.5[(1+\beta) \times \text{Parent}^1 - (1-\beta) \times \text{Parent}^2]$$
$$\text{Child}^2 = 0.5[(1-\beta) \times \text{Parent}^1 + (1+\beta) \times \text{Parent}^2],$$

where η should be a non-negative real number. Large values of η increase the probability of creating descendants that are close to its parents, while small values of η allow distant solutions to be selected as children. In this work we adopt $\eta = 3.0$, as it provides a good compromise between these two extremes.

When applied to a gene, mutation changes the value of the parameter stored at that location. Figure 2(d) exemplifies this operation. The parameter encoded in the second gene $\left(Z_2^1\right)$ is mutated to a new value $\left(Z_2^m\right)$. Two different operators were tested in this work: random mutation and sigma mutation. They differ in how they perform the modification. The first operator randomly selects a new value on the corresponding interval. As for sigma mutation, it obtains the new value for the fitting parameter according to the following expression:

$$Z_2^m = Z_2^1 + \sigma \times \left(Z_2^{\max} - Z_2^{\min}\right) \times N(0,1), \tag{2}$$

where $N(0, 1)$ represents a random value sampled from a standard normal distribution, $Z_2^{\max}$ and $Z_2^{\min}$ represent the upper and lower bounds of the interval for the fitting parameter undergoing mutation and σ is a parameter from the algorithm. The performance of different crossover and mutation operators, including an analysis of the influence of its application rates, is studied in subsection 3.1.

The new generation is formed as soon as the genetic operators finish their work. Individuals are then locally optimized and evaluated. Afterwards, a straightforward elitist strategy ensures that the quality of the best individual never decreases along the optimization. The described process is repeated until a termination criterion is met. In our study, we let the GA run until the number of calls to the evaluation function reaches a pre-specified number.

The best individual (i.e., the best set of fitting parameters) obtained within a given number of runs is considered as the best solution found by the described GA.

2.2. Direct fitting procedure

The procedure to obtain directly an analytic potential energy curve $V(R)$ of a particular electronic state of a diatomic molecule consists to perform a nonlinear least-squares fit by using the available experimental information, such as the vibrational spectrum, and occasionally theoretical data (e.g., *ab initio* electronic energies). In the present paper, we consider a two-step GA to perform this task: first, the GA is applied to estimate the appropriate interval for each one of the fitting parameters by employing only the *ab initio* points in the fit; subsequently, the GA is used to find the best set of potential parameters from the appropriate intervals by fitting both experimental spectroscopic and *ab initio* data. In such a case, the objective function (equation (1)) should be rewritten as

$$\chi^2(\mathbf{z}) = \sum_{j=1}^{N_{\text{exp}}} \omega_j^{\text{exp}} \left[E_j^{\text{cal}}(\mathbf{z}) - E_j^{\text{exp}}\right]^2 + \sum_{j=1}^{N_{ab}} \omega_j^{ab} [V^{\text{fit}}(R_j; \mathbf{z}) - V^{ab}(R_j)]^2, \tag{3}$$

where N_{exp} and N_{ab} are, respectively, the numbers of vibrational energies and *ab initio* points, $E_j^{\text{cal}}(\mathbf{z})$ is the jth theoretical vibrational energy calculated for the adjustable potential function $V^{\text{fit}}(R; \mathbf{z})$, E_j^{exp} is the jth experimental vibrational energy level, $V^{ab}(R_j)$ is the *ab initio* energy associated with the distance R_j and $\{\omega_j\}$ are the weights given to each one of known experimental and theoretical data values.

The analytic potential function applied here is the simple and well-known extended-Rydberg form, i.e.,

$$V(R) = -D_{\text{e}} \left(1 + \sum_{k=1}^{p} a_k \rho^k\right) \exp(-a_1 \rho), \tag{4}$$

where $\rho = R - R_{\text{e}}$ (R_{e} is the equilibrium geometry); note that for $R = R_{\text{e}}$ the potential becomes equal to $-D_{\text{e}}$, i.e., the symmetric of the dissociation energy. Within this model, we assume all the $a_k (k = 1, \ldots, p)$, D_{e} and R_{e} as fitting parameters that characterize the individual with $n = p + 2$ genes in GA, i.e. $\mathbf{z} = (z_1 = a_1, \ldots, z_p = a_p, z_{p+1} = D_{\text{e}}, z_{p+2} = R_{\text{e}})$. (Although many other functional forms could be applied in alternative to equation (4), this has been chosen for simplicity reasons.) On the other hand, the vibrational eigenenergies $E_j^{\text{cal}}(\mathbf{z})$ are solutions of the time-independent Schrödinger equation

$$\left[-\frac{\hbar^2}{2\mu}\frac{\mathrm{d}^2}{\mathrm{d}R^2} + V^{\text{fit}}(R; \mathbf{z})\right] \psi_j(R) = E_j^{\text{cal}}(\mathbf{z}) \psi_j(R) \tag{5}$$

obtained using the discrete variable representation (DVR) method with an equally spaced grid [80, 81]. In equation (5), μ designates the reduced mass of the diatomic system, while $\psi_j(R)$ is the j-state vibrational wavefunction.

3. Results and discussion

3.1. Performance of the GA

Since the main goal of the present work is the development of an efficient GA to be applied in the direct fit of potential energy curves to both *ab initio* and spectroscopic data, we have made some preliminary tests to obtain a better understanding of the algorithm's behaviour. Specifically, we need to know which genetic operators (from those described in section 2.1) perform best, in what concerns both the convergence rate

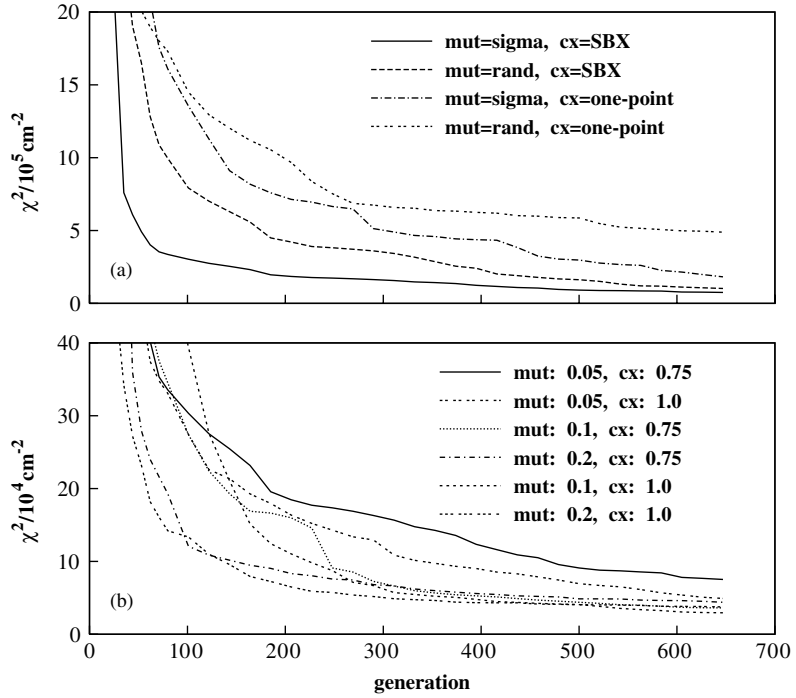

Figure 3. Influence of the crossover and mutation operators in the performance of the GA: (a) tests for combinations of different operators; (b) tests for different rates of application of the SBX crossover and sigma mutation. In the insert caption 'mut' refers to mutation, while 'cx' designates crossover; the numbers are crossover and mutation rates (see the text).

of the GA and the best solution found (i.e., the minimum objective-function achievement within the least-squares fitting procedure). Additionally, we also tested the influence of different combinations of crossover and mutation rates. All the test sets have been carried out by applying the GA for fitting both *ab initio* points and experimental vibrational levels of NaLi to the nine-parameter extended-Rydberg function (see subsection 3.3 for details).

For these initial tests, the settings of the GA are the following: number of runs, 10; objective-function evaluations, 8×10^3; population size, 12; tourney size, 2; crossover operators, {One-point, SBX}; mutation operators, {Random, Sigma}; σ, 0.1; crossover rate, {0.75; 1.0}; mutation rate, {0.05, 0.1, 0.2}; LSL, 1000. The number of objective-function evaluations is specified in such a way that one can extract some information about the performance of GA with a minimum computational cost. The main results are displayed in figure 3 and tables 1 and 2.

The evolution of the mean best fitness (MBF) (i.e., the average of the best fitness values over all runs) along the generations is shown in figure 3(a); the corresponding values for the MBF and best-ever-fitness at the end of the optimization are given in table 1. These tests were performed with a crossover rate of 0.75 and a mutation rate of 0.05.

It is clear from figure 3(a) and table 1 that the best results are obtained when SBX is used in combination with sigma mutation. On the other hand, the worst solutions occur when one-point crossover and random mutation are used.

Moreover, the results show that the influence of crossover for discovering good quality solutions is higher than that of

Table 1. Tests performed for different types of crossover and mutation operators; the crossover rate is 0.75, while the mutation rate is 0.05. The values of the χ^2 objective-function presented in the table[a] are in cm^{-2}.

Crossover operators	Mutation operators	
	Random	Sigma
One-point	488 431	182 308
	21 262	20 772
SBX	102 616	75 146
	17 329	17 122

[a] The mean best fitness (MBF) and the best-ever-fitness values of the χ^2 objective function are given in first and second entries, respectively.

Table 2. Tests performed for different rates of the SBX crossover and sigma mutation operators. The values of the χ^2 objective-function presented in the table[a] are in cm^{-2}.

Crossover rate	Mutation rate		
	0.05	0.1	0.2
0.75	75 146	36 120	43 861
	17 122	14 360	26 013
1.0	49 201	29 469	37 581
	13 831	13 394	21 635

[a] The mean best fitness (MBF) and the best-ever-fitness values of the χ^2 objective function are given in first and second entries, respectively.

mutation. Switching from one-point crossover to SXB (while maintaining the same mutation operator) clearly improves both the best solution found and the MBF. In contrast, changing from random to sigma mutation (while maintaining the same crossover operator) leads only to a small improvement in the achieved results.

It is important to note that these results should be considered as preliminary, as they were obtained after a limited number of evaluations. In depth conclusions about the performance of different genetic operators would need more generations, since the slope of the lines in figure 3(a) suggests that the GA did not yet reach the end of the optimization process when the termination criterion is met. In spite of that, the results show a trend that favours the combination of SBX and sigma mutation as the best set of operators. We have then performed extra tests to fine-tune the crossover and mutation rate parameters. The evolution of the MBF along the generations for various sets of parameters is represented in figure 3(b), while the corresponding values for the MBF and best-ever-fitness at the end of the optimization are given in table 2. For comparison purposes, we repeat in panel (b) the solid curve for SBX and sigma mutation (now designated as 'mut: 0.05, cx: 0.75') from panel (a).

It is clear that the improvement arising from different combinations of parameter rates appear to be less relevant than that from the different sets of crossover and mutation operators in panel (a). (Note that the χ^2-scale of panel (b) has been reduced by about one order of magnitude in relation to panel (a).) Nevertheless, it is apparent from figure 3(b) and table 2 that all combinations of crossover and mutation rates tend to perform better than both the standard set 'mut: 0.05, cx: 0.75'; the MBF is lower in all cases, whilst the best-ever-fitness is higher only in tests performed with a mutation rate of 0.2. The results suggest that a somewhat intensive application of genetic operators helps the GA to efficiently explore the search space (see also below). This might be related to the existence of a local optimization method. When applied, L-BFGS drives the search towards the nearest local minimum from where it departs. A vigorous shake provided by the genetic operators is then needed to allow the GA to escape from these local optima. Nonetheless, the results also show that a mutation rate of 0.2 is excessive. When this value is adopted, the GA has difficulties in focusing in promising areas of the search space. The evolution of the MBF presented in figure 3(b) confirms this supposition: in the beginning of the optimization, trials performed with a mutation rate of 0.2 perform better, but over time they become worse than tests performed with a mutation rate of 0.1 showing that they are not able to perform a more localized search. As a consequence of this analysis, we have decided to use the parameters from the set 'mut: 0.1, cx: 1.0' in the 'real' fitting procedures, whose results are presented in subsections 3.2 and 3.3.

In addition, we have also tested the influence of the population size in the behaviour of the algorithm. Specifically, we performed an additional test where we have increased the population size by a factor of 5, while reducing the number of evaluations of the objective function by the same factor. The result was clearly worse than with the population of 12 individuals, so this size has been used in the tests whose results are described in subsections 3.2 and 3.3.

To conclude the preliminary analysis, we have also calculated the best-ever-fitness obtained when all genetic operators are 'switched off' while keeping the local optimization during the same number of objective-function evaluations (i.e., 8×10^4 which corresponds to 10 runs); in this case, the value of the best-ever-fitness is 16299 cm^{-2}. This result is clearly worse than the best solution found by the hybrid algorithm, confirming the importance of the global exploration performed by the genetic component of the optimization method in the context of data fitting problems.

3.2. Ar_2 potential

We begin to apply the fitting procedure suggested in this work to the ground-electronic state of the Ar_2 diatomic. Since Ar_2 is a van der Waals molecule with a small energy well, the number of vibrational levels involved in the fitting procedure of this system is only six (or five consecutive vibrational energy differences). Thus, this is probably less demanding than the case of NaLi (subsection 3.3) for applying the GA fitting procedure. Nevertheless, Ar_2 is a benchmark van der Waals diatomic studied in many contexts (namely for testing a new fitting procedure [6]) for which very accurate CCSD(T) *ab initio* data are available [47, 48]; also, the energy differences between consecutive vibrational levels have been extracted from the measured vibronic spectra of Ar_2 [49].

As mentioned in section 2, we have applied the GA in a two-step fitting procedure to the model function of equation (4); for Ar_2, we have carried out fits (in a total of three sets) considering 5, 7 and 9 fitting parameters in equation (4). In the first step, we have fitted 19 CCSD(T) *ab initio* points (beginning from $R = 3.2$ Å up to $R = 20$ Å, which correspond to interaction energies of 170.353 cm^{-1} and -0.0050 cm^{-1}, respectively) published by Patkowski *et al* [48]. Note that, among all the CCSD(T) points of [48], only those in the attractive region of the potential are important to adequately describe the spectroscopy of Ar_2, and hence we have not used their data from the high repulsive region. With these points, the GA converges to the putative best value of χ^2 after a small number of generations, and values of the fitting parameters so obtained give us insight about the searching intervals to be used in the second step of the fit. Then, in the second step, the GA is applied to fit both the previous *ab initio* points and the five consecutive vibrational energy differences obtained from the Ar_2 vibronic spectrum of Herman *et al* [49]. Since the main goal is to accurately fit the experimental spectrum data, we have arbitrarily assigned a weight of 100 to the five consecutive vibrational energy differences, while keeping a weight of 1 for the *ab initio* points. The GA is allowed to run during, at least, 2×10^4 evaluations of the χ^2 function, which guarantees the convergence of the method. The best set of parameters has been selected among four different runs of the GA, and the numeric values are given in table 3 for the three fits considered for Ar_2.

Moreover, we represent in table 3 the root-mean-square deviation (rmsd) values between the fitted curves of Ar_2 and

Table 3. Parameters and root-mean-square deviation (rmsd) for Ar_2 obtained in three fits of the extended-Rydberg function to *ab initio* and spectroscopic data.

	Fits[a]		
	1	2	3
	Parameters[b]		
D_e	4.523 2064(−4)	4.527 0502(−4)	4.525 2211(−4)
R_e	7.123 8745	7.119 8023	7.121 2913
a_1	1.105 4902	1.054 4889	1.054 0163
a_2	−0.196 698 44	−0.243 200 66	−0.239 926 00
a_3	8.172 112 57(−2)	0.100 331 11	9.554 7537(−2)
a_4		−1.214 0720(−2)	−1.367 8454(−2)
a_5		1.625 8235(−3)	4.182 6422(−3)
a_6			−7.463 6184(−4)
a_7			6.427 7258(−5)
	rmsd[c]		
Ab initio	0.447 (0.272)	0.134 (0.043)	0.107 (0.017)
Vibrational	0.090 (0.191)	0.019 (0.067)	0.008 (0.069)

[a] Fits 1, 2 and 3 include 5, 7 and 9 fitting parameters in equation (4), respectively.
[b] Values are in atomic units.
[c] Values are in cm^{-1}; values in parenthesis are for the fit of *ab initio* data alone.

both the *ab initio* points and the vibrational energy differences; values in parenthesis are for the fits of the *ab initio* points alone (i.e., first step of the GA fitting procedure). As expected, the rmsd decreases with increasing number of fitting parameters, which is more significant on passing from 5 (fit 1) to 7 (fit 2). Of course, this is not always true for the spectroscopic data when fitting only the *ab initio* points (e.g., the rmsd increases from 0.067 cm^{-1} in fit 2 to 0.069 cm^{-1} in fit 3). However, the rmsd of the vibrational energy differences clearly improves after applying the second step of the GA fitting, especially for fit 1, which is achieved to the expenses of worsening the corresponding values for the *ab initio* data.

It is worth noting that with only five fitting parameters (fit 1) one obtains already a good fit of the spectroscopic data, and even the fit of the *ab initio* points alone leads to a rmsd ∼0.2 cm^{-1} for the vibrational energy differences. This good agreement between experiment [49] and fitting model results from the fact that one starts from very accurate CCSD(T) *ab initio* data [48]. Indeed, Patkowski *et al* [48] have calculated vibrational energy differences close to the experimental values [49] by using an analytic function proposed by Korona *et al* [82] to which their CCSD(T) data have been fitted. This potential [48] is compared with the one arising from our best fit (i.e., fit 3) in the bottom panel of figure 4, where we represent the spectroscopic-relevant part of the Ar_2 potential. Since the two curves are essentially coincident within the scale of the figure, we also represent in the top panel of figure 4 the difference between the two potentials. It is shown in this figure that, except for internuclear distances smaller than the equilibrium geometry where the major difference can reach ∼2 cm^{-1}, the two potential curves differ by no more than 0.4 cm^{-1} in the spectroscopic-relevant region; the equilibrium geometry and the dissociation energy from our nine-parameter potential are, respectively, 7.121a_0 and 99.32 cm^{-1} (which corresponds to $D_e = 4.525 \times 10^{-4} E_h$, as shown in table 3), while the corresponding values for the potential of Patkowski *et al* [48] are 7.119a_0 and 99.27 cm^{-1}. However, these discrepancies are sufficient to bring the computed vibrational data from our potential to coincide (within the error bars)

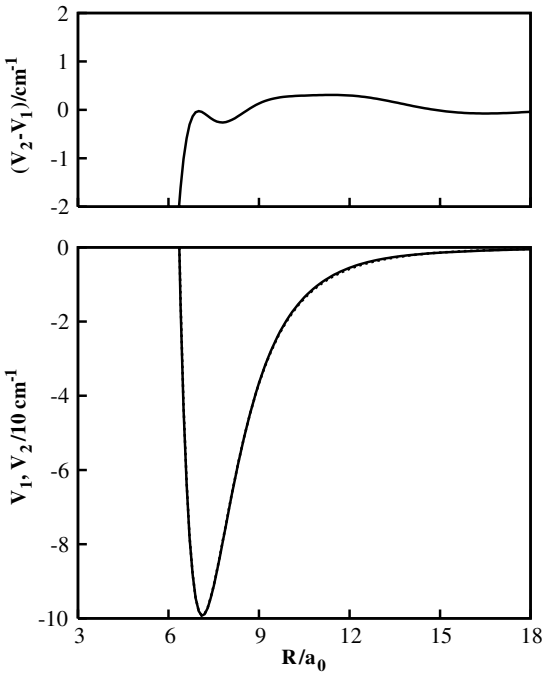

Figure 4. Potential energy curve for the Ar_2 dimer obtained by joint fitting the *ab initio* and spectroscopic data (V_2, solid line) to a nine-parameter potential model given by equation (4); parameters are from fit 3 in table 3. Although essentially coincident with V_2 (within the scale of the figure), it also displayed the curve (V_1, dotted line) proposed by Patkowski *et al* [48], which results from a least-squares fit of their own *ab initio* data to a different model function [82]. Also shown in the top panel is the difference between the two curves (i.e., $V_2 - V_1$).

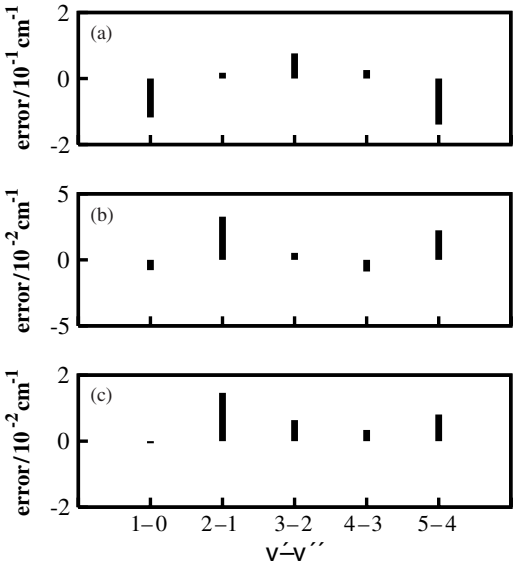

Figure 5. Errors in the energy differences between consecutive Ar_2 vibrational levels (in relation to the corresponding experimental values) predicted by the potential curves obtained in fit 1 (panel (a)), fit 2 (panel (b)) and fit 3 (panel (c)) of table 3. Note the differences in the energy scales of the panels.

Table 4. Energy differences[a] between consecutive vibrational levels (with rotational quantum number $J = 0$) of Ar_2 predicted by experiment and different potential models (including the one from the GA fit 3 of table 3).

$v' - v''$	(this work) E^{cal}	E^{cal} ([48])	E^{cal} ([83])	E^{exp} ([49])
$1-0$	25.69	25.75	25.68	25.69 ± 0.01
$2-1$	20.57	20.48	20.56	20.58 ± 0.02
$3-2$	15.57	15.44	15.58	15.58 ± 0.02
$4-3$	10.91	10.79	10.92	10.91 ± 0.03
$5-4$	6.83	6.76	6.83	6.84 ± 0.07

[a] All values are in cm^{-1}.

with experiment [49], while the potential of Patkowski *et al* [48] presents small disagreement for some vibrational energy differences, as displayed in table 4. It is apparent from table 4 that our potential is spectroscopically as good as the one from Aziz [83], which is considered as the benchmark for Ar_2. Recently, a potential for Ar_2 has been published [6] which is probably as good as those in the spectroscopic-relevant region.

The evolution of the error in vibrational energy differences ($v' - v''$) with the number of parameters included in the fitting function of equation (4) is shown in figure 5; note that we have calculated the error having as reference the corresponding experimental values of Herman *et al* [49]. It is interesting to observe in figure 5 that the error in all $v' - v''$ values diminishes as the number of fitting parameters increase from 5 (panel (a)) up to 9 (panel (b)). While the major errors for fit 1 (panel (a)) are in differences $1 - 0$ and $5 - 4$, they are in differences $2 - 1$ and $5 - 4$ for fit 2 (panel (b)) and fit 3 (panel (c)). Such errors for fit 3 are, however, below the experimental uncertainty (see table 4); even for fit 2, only the $2 - 1$ (with an error of 0.033 cm^{-1}) is still larger than the experimental uncertainty.

Table 5. MCQDPT *ab initio* points used for fitting the NaLi($X^1\Sigma^+$) diatomic potential in this work (see the text).

R/a_0	Energy/$10^{-6}E_h$
3.5	178 04
3.75	331 8
4.0	−806 2
4.25	−167 46
4.5	−231 15
5.0	−302 92
5.5	−321 77
6.0	−307 67
6.5	−275 25
7.0	−234 45
7.5	−191 89
8.0	−151 78
8.5	−116 57
9.0	−873 5
10.0	−463 4
11.0	−236 2
12.0	−119 4
13.5	−445
15.0	−184
17.0	−69
20.0	−14
30.0	−1

3.3. NaLi potential

A more challenging test for the GA fitting procedure proposed in this work is the NaLi diatomic system. Different from Ar_2, the electronic ground state of NaLi presents a deep well with 48 predicted [38] vibrational levels; the 44 lowest lying of these vibrational term values (or levels) of NaLi($X^1\Sigma^+$) have been studied experimentally by one of us [38] through the application of a Fourier transform spectroscopy and laser-induced fluorescence technique. An extra difficulty arises in this case, because the set of *ab initio* points we have used in the fit of the electronic ground state of NaLi is not as accurate as that reported above for Ar_2. Indeed, we have used in the fit a set of 22 MCQDPT *ab initio* points obtained from a state-average complete active-space self-consistent field (SA-CASSCF) involving five singlet electronic states using the aug-cc-pCVQZ atomic basis set of Dunning and collaborators [84–86]. These *ab initio* points arise from a preliminary study [31] on the radiative transitions of the NaLi molecule and are far from being the most accurate for this system (see [31]); nevertheless, we have used them to impose a greater challenge for the present methodology. In turn, this set of *ab initio* points covers a range that extends from $R = 3.5\,a_0$ up to $R = 30\,a_0$, and the corresponding numerical values are given in table 5 for completeness.

As before, the first step involves the fit of the extended-Rydberg function (equation (4)) to the *ab initio* data alone, so that one may have insight on the best searching intervals for parameters to be used in the second step. We allow the GA to proceed up to 1×10^5 evaluations of the objective function, χ^2. In turn, the second step takes the relevant information

Table 6. Parameters and root-mean-square deviation (rmsd) for NaLi obtained in four fits of the extended-Rydberg function to *ab initio* and spectroscopic data.

	Fits[a]			
	1	2	3	4
		Parameters[b]		
D_e	0.032 413 373	0.032 272 999	0.032 399 123	−0.032 365 540
R_e	5.500 3617	5.513 9121	5.498 2039	5.486 5604
a_1	1.057 2296	0.877 4290	1.095 2386	0.985 288 74
a_2	0.349 386 79	0.178 001 40	0.393 744 55	0.278 443 97
a_3	4.382 6194 (−2)	−4.739 1504 (−3)	6.361 0957 (−2)	3.001 0578 (−2)
a_4		−4.588 0738 (−3)	2.560 8268 (−3)	−1.031 6092 (−3)
a_5		4.231 5415 (−4)	−1.075 6819 (−3)	−1.891 5786 (−3)
a_6			−1.164 4694 (−4)	−8.637 6274 (−5)
a_7			4.233 0714 (−5)	1.354 5263 (−4)
a_8				−1.761 8077 (−5)
a_9				7.215 8300 (−7)
		rmsd[c]		
ab initio	30.792 (18.058)	44.265 (8.489)	22.327 (0.941)	22.634 (0.733)
vibrational	10.474 (28.181)	2.265 (36.631)	0.573 (47.221)	0.184 (38.409)

[a] Fits 1, 2, 3 and 4 include 5, 7, 9 and 11 fitting parameters in equation (4), respectively.
[b] Values are in atomic units.
[c] Values are in cm^{-1}; values in parenthesis are for the fit of *ab initio* data alone.

from previous step to perform a joint fit of the *ab initio* and spectroscopic data sets; a weight of 1 is attributed to the *ab initio* points, while the corresponding value for vibrational levels was 100. This two-step procedure has been applied in four fits of the extended-Rydberg function (equation (4)), where 5 (fit 1), 7 (fit 2), 9 (fit 3) and 11 (fit 4) fitting parameters have been included. In each case, we have carried out a total of 4 runs of the GA with 2×10^4 or 3×10^4 (for fit 4) evaluations of χ^2-function per run, so that one guarantees the convergence of the GA fitting. For long runs, however, we added a straightforward hyper-mutation functionality to the GA, which is applied to all genes of the individuals from the current population whenever the best solution is not improved for 50 generations; the goal is to guarantee that the search does not get stuck on a local optimum. The best set of fitting parameters, so obtained, are shown in table 6 for fits 1–4.

Since the parameters R_e and D_e were not imposed *a priori*, it would be interesting to compare their experimental values [37, 38] (i.e., $R_e = 5.4594 \pm 0.0038\,a_0$ and $D_e = 7105.5 \pm 1.0\ cm^{-1} \approx 0.032\,375\,041 \pm 0.000\,004\,56 E_h$) with the corresponding ones given in table 6 for each fit. We observe that, for both R_e and D_e, the best agreement with experiment arise in the case of fit 4, being the errors $-0.0271\,a_0$ and $2.085\ cm^{-1}$, respectively; the corresponding errors for R_e (D_e) arising in fit 1, fit 2 and fit 3 are $-0.0409\,a_0$ ($-8.413\ cm^{-1}$), $-0.0545\,a_0$ ($22.396\ cm^{-1}$) and $-0.0388\,a_0$ ($-5.285\ cm^{-1}$), respectively. These errors may be attributed to an insufficient flexibility of the model function used in the fit. Curiously, fit 1 gives better agreement with experiment for both R_e and D_e than fit 2.

It is also represented in table 6 the rmsd values between the fitted curves for NaLi and both the *ab initio* points and the experimental vibrational levels; values in parenthesis are for the fits of the *ab initio* points alone (i.e., first step of the GA fitting procedure). From a comparison between tables 6 and 3, we note that the rmsd values for NaLi are always larger than the corresponding ones for Ar_2. Indeed, as stated above, NaLi is a much more difficult case, because it has a larger number of vibrational levels to be fitted than Ar_2 and less accurate *ab initio* points were employed in the present fit.

It is shown in table 6 that increasing the number of fitting parameters when the *ab initio* points are fitted alone clearly improves the corresponding value of the rmsd, while generally worsening the rmsd for the vibrational levels (that are not being fitted); an exception occurs when passing from 9 (fit 3) to 11 (fit 4) fitting parameters: the rmsd for the vibrational levels diminishes from 47.221 cm^{-1} to 38.409 cm^{-1}. In contrast, the *ab initio* rmsd diminishes from 18.058 cm^{-1} for the five-parameters fit to 0.733 cm^{-1} for the eleven-parameters fit. On the other side, when performing the joint fit of *ab initio* and spectroscopic data (corresponding to the second step of our method), the rmsd of the vibrational levels may be reduced up to 0.184 cm^{-1} (fit 4), while the best rmsd for the *ab initio* points is as high as 22.327 cm^{-1} (fit 3). In fact, the impossibility of having simultaneously low rmsd values for both *ab initio* and vibrational levels is a clear indication of a certain lack of accuracy of the theoretical data.

The errors in the vibrational levels for NaLi are displayed in figure 6 for fit 1 (panel (a)), fit 2 (panel (b)), fit 3 (panel (c)) and fit 4 (panel (d)). It is apparent from this figure that the error in the levels strongly diminishes as the number of fitting parameters of the model function increases. Indeed, one obtains errors of more than 20 cm^{-1} for fit 1, which is reduced for less than 5 cm^{-1} for fit 2. In turn, fit 3 leads to errors that are all below 1 cm^{-1} (but some are very near this value), while in fit 4 all errors are below 0.5 cm^{-1}. In addition, most of the errors in the levels of fit 4 are even below 0.2 cm^{-1}, being the exception the set of levels from $v = 24$ up to $v = 33$. Furthermore, it is interesting to note that such sets of levels appear to have the major errors in all fits.

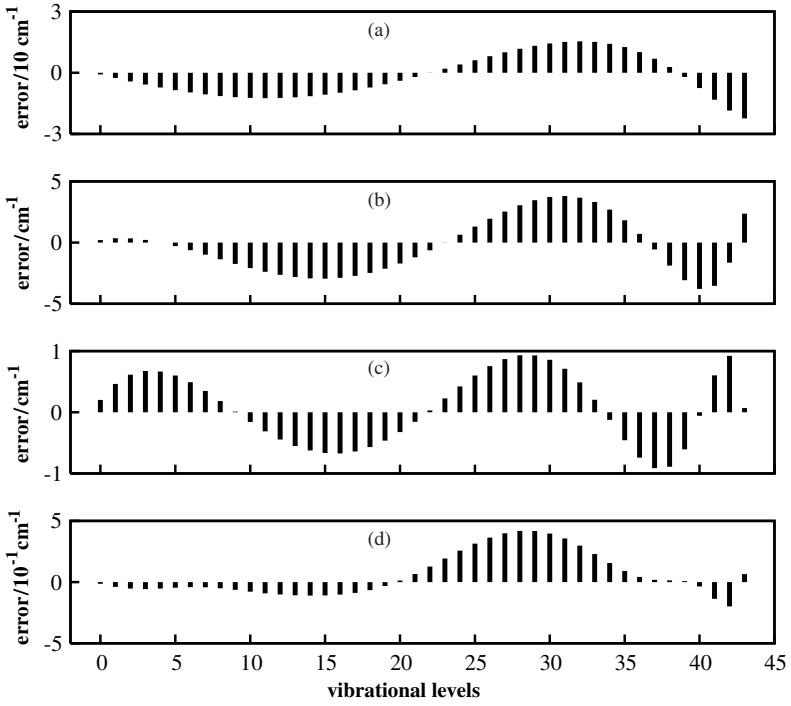

Figure 6. Errors in the NaLi vibrational levels (in relation to the corresponding experimental values) predicted by the potential curves obtained in fit 1 (panel (a)), fit 2 (panel (b)), fit 3 (panel (c)) and fit 4 (panel (d)) of table 6. Note the differences in the energy scales of the panels.

We represent in table 7 the experimental vibrational levels of NaLi as well as the corresponding ones obtained with the diatomic potential arising from fit 4; values in parenthesis correspond to fit 4, but for fitting the *ab initio* data alone. It is patent in this table that, after the two-step procedure, the major error in the vibrational levels arises for $v = 28$ (with $\Delta E = 4.180 \times 10^{-1}$ cm^{-1}). Conversely, for fitting the *ab initio* points alone, the errors in the vibrational levels extend from 1.771 cm^{-1} for $v = 0$ up to 62.825 cm^{-1} for $v = 30$. In addition, we compare in figure 7 (bottom panel) the two potentials curves obtained in fit 4 from fitting both *ab initio* and spectroscopic data, and *ab initio* points alone. Although the two curves essentially overlap within the scale of the figure, it is clear from the top panel of figure 7 that the difference between the two potentials may reach values close to ± 50 cm^{-1} in the spectroscopic-relevant region. This shows the extension of the correction in the potential model imposed by the joint fit of inaccurate *ab initio* data and the experimental vibrational levels of the electronic ground state of the NaLi molecule.

Another interesting outcome is related to the fact that the NaLi($X^1\Sigma^+$) potential obtained in fit 4 has 48 vibrational levels, as predicted by Fellows [38] from his experimental results on this system; the four highest vibrational levels calculated with potential from fit 4 (not shown in table 7) have the following values (in cm^{-1}): 7086.385, 7094.272, 7099.603 and 7102.938 for $v = 44, 45, 46$ and 47, respectively. In contrast, the diatomic curves from fit 1 and fit 2 present only 45 vibrational levels, while one calculates bound states up to $v = 46$ (i.e., 47 levels) for the potential obtained from fit 3.

Finally, it would be interesting to apply the present methodology for fitting spectroscopic data alone, once such experimental information is in general more accurate than *ab initio* points. We have used the GA to fit the 44 vibrational levels of NaLi to the eleven-parameter extended-Rydberg function (i.e., similar to fit 4). In this case, the rmsd of the vibrational levels has been reduced to 0.140 cm^{-1}, which constitutes a significant improvement in relation to the above-mentioned fit 4, i.e., 0.184 cm^{-1} (that includes both *ab initio* and experimental data). However, the decrease in the rmsd is achieved at the expense of a certain deterioration of the global behaviour of the potential. Indeed, the fit of vibrational levels alone tends to reduce the width and the well depth of the potential which leads it to present only 44 vibrational levels (i.e., the same number as the fitted ones), in contrast to the experimentally estimated [38] value of 48. This shortcome of the arising potential may be attributed to the simple model function (with reduced extrapolation ability) used here; hence, in the absence of *ab initio* information, a more accurate description of the spectroscopic properties (within this model) can be only guaranteed after including a larger set of vibrational–rotational data in the direct fitting procedure. Thus, the role of the *ab initio* points (extending for the large range $3.5 \leqslant R/a_0 \leqslant 30$) in fit 4 appears to be the enlargement of the potential well, leading to a deeper potential, which increases the number of vibrational levels presented by the potential curve so obtained.

Table 7. Comparison between the vibrational energies[a] (in cm^{-1}) for NaLi obtained from fit 4 (E^{cal}) of table 6 and the corresponding experimental values (E^{exp}) used in the fit.

Vibrational level	$E^{cal}(cm^{-1})$	E^{exp} (cm^{-1})	$\Delta E^a = (E^{exp} - E^{cal})$ (cm^{-1})
0	127.846 (126.061)	127.832	−0.014 (1.771)
1	381.162 (375.918)	381.122	−0.040 (5.204)
2	631.167 (622.619)	631.115	−0.052 (8.496)
3	877.845 (866.128)	877.788	−0.057 (11.660)
4	1121.170 (1106.408)	1121.117	−0.053 (14.709)
5	1361.114 (1343.417)	1361.068	−0.046 (17.651)
6	1597.644 (1577.111)	1597.603	−0.041 (20.492)
7	1830.720 (1807.442)	1830.678	−0.042 (23.236)
8	2060.301 (2034.361)	2060.250	−0.051 (25.889)
9	2286.337 (2257.813)	2286.274	−0.063 (28.461)
10	2508.777 (2477.741)	2508.699	−0.078 (30.958)
11	2727.563 (2694.084)	2727.472	−0.091 (33.388)
12	2942.633 (2906.777)	2942.532	−0.101 (35.754)
13	3153.920 (3115.753)	3153.812	−0.108 (38.059)
14	3361.349 (3320.939)	3361.239	−0.110 (40.300)
15	3564.844 (3522.257)	3564.735	−0.109 (42.478)
16	3764.320 (3719.627)	3764.218	−0.102 (44.591)
17	3959.687 (3912.961)	3959.599	−0.088 (46.638)
18	4150.848 (4102.169)	4150.783	−0.065 (48.614)
19	4337.700 (4287.151)	4337.668	−0.032 (50.517)
20	4520.130 (4467.806)	4520.143	0.013 (52.337)
21	4698.021 (4644.021)	4698.088	0.067 (54.067)
22	4871.245 (4815.680)	4871.372	0.127 (55.692)
23	5039.663 (4982.656)	5039.856	0.193 (57.200)
24	5203.130 (5144.815)	5203.386	0.256 (58.570)
25	5361.486 (5302.013)	5361.800	0.314 (59.786)
26	5514.561 (5454.096)	5514.924	0.363 (60.828)
27	5662.171 (5600.895)	5662.570	0.399 (61.675)
28	5804.119 (5742.232)	5804.537	0.418 (62.305)
29	5940.190 (5877.913)	5940.608	0.418 (62.695)
30	6070.154 (6007.726)	6070.551	0.397 (62.825)
31	6193.760 (6131.444)	6194.117	0.357 (62.673)
32	6310.741 (6248.819)	6311.039	0.298 (62.220)
33	6420.806 (6359.578)	6421.035	0.230 (61.457)
34	6523.645 (6463.428)	6523.801	0.156 (60.373)
35	6618.931 (6560.047)	6619.022	0.091 (58.974)
36	6706.321 (6649.088)	6706.364	0.043 (57.276)
37	6785.467 (6730.178)	6785.485	0.018 (55.307)
38	6856.028 (6802.925)	6856.042	0.014 (53.117)
39	6917.697 (6866.933)	6917.706	0.008 (50.773)
40	6970.228 (6921.836)	6970.192	−0.036 (48.356)
41	7013.454 (6967.348)	7013.318	−0.136 (45.970)
42	7047.278 (7003.353)	7047.079	−0.199 (43.726)
43	7071.578 (7029.984)	7071.644	0.066 (41.660)

[a] Values in parenthesis refer to the fit of the *ab initio* data alone (i.e., first step of the GA fitting).

4. Conclusions and final remarks

We have proposed a two-step GA fitting procedure to obtain diatomic potential curves with spectroscopic accuracy. The most relevant parameters of the GA have been tuned to enhance the performance by doing some tests prior to the 'real' applications. Then, the method was applied to Ar_2, which is a benchmark van der Waals system, and to the more challenging NaLi molecule. Specifically, we have fitted the extended-Rydberg function (equation (4)) (by considering various sets of fitting parameters) to both *ab initio* and spectroscopic data. The method has shown to be robust for searching the best parameters that fit the *ab initio* and spectroscopic data. In the case of NaLi, the application of the present methodology to fit the vibrational data alone (i.e., without including the *ab initio* points) led to a smaller rmsd value, but the potential shows worse global behaviour than the corresponding one obtained from the joint fitting procedure.

Although it has been tested here by applying the extended-Rydberg function, this methodology may be extended to other analytic diatomic potentials. Moreover, the application of the two-step GA method to fit diatomic potential energy curves directly to the experimental wave numbers resulting from the individual transitions between the rovibrational levels of

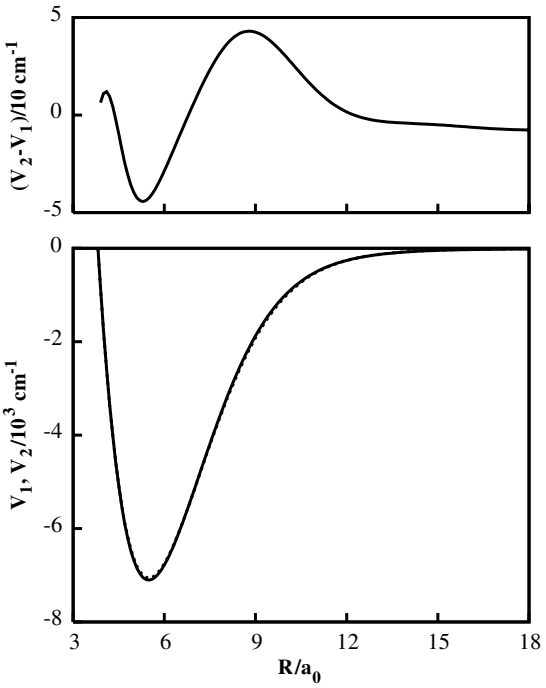

Figure 7. Potential energy curves for the electronic ground state of NaLi obtained by joint fitting the *ab initio* and spectroscopic data (V_2, solid line), and by fitting the *ab initio* energies alone (V_1, dotted line); in both cases, 11 fitting parameters have been used in the model potential of equation (4). Also shown in the top panel is the difference $V_2 - V_1$.

the electronic excited state and the corresponding ones of the electronic ground state is straightforward and it will be employed in the NaLi and RbCs systems in the near future.

Finally, we should comment on the possibility of extending the present methodology to obtain analytic potential energy functions for polyatomic systems. Although the results of the present work for diatomic molecules are encouraging in the perspective of applying GAs for fitting multidimensional PESs, it should be pointed out that it is a very difficult task comprising many problems. For example, a PES to be used in dynamics scattering calculations must correctly incorporate the relevant asymptotic dissociation limits, while studies focusing on the vibrational spectrum need a good representation of the potential well. Both cases oblige for an adequate choice of the functional form that has to reproduce accurately the available data for the system. It should be emphasized that a large amount of data is necessary to obtain a good description of a multidimensional PES and, hence, it is usually imperative to combine experimental information with *ab initio* energies in the fitting procedure. In turn, a common approximation to build up potential energy surfaces is based on the many-body expansion (MBE) method [87], where the energy is partitioned in the corresponding two-body, three-body, etc interaction terms, being each n-body component represented by a specific function; by construction, it guarantees the correct description of the asymptotic dissociation limits. Thus, the analytic potential is a sum of all diatomic functions (e.g., extended-Rydberg), the three-body components and higher-order terms (when applied). The functional form for MBE terms other than the diatomic ones is not so easy to establish, since it involves various coordinates (i.e., it is a multidimensional function). An alternative approach for such a functional representation consists of using neural networks [88], which have been successfully applied in the construction of PESs [89–94]. In spite of using the MBE approach [87], one may simply build up a mode-expanded PES to obtain a good representation of the spectroscopic relevant region of the configuration space. This approximation consists of a linear combination of n-mode cuts of the PES (i.e., sections of the PES for which only n normal coordinates are different from zero) and it has been successfully applied by Christiansen and collaborators [95, 96]. Concerning the fitting of spectroscopic data (to represent accurately the potential well), we can still use the GA providing that one applies an efficient method to calculate the vibrational states (e.g., to substitute the DVR method used in this work for diatomics). Recently, Christiansen [97, 98] has proposed a vibrational-coupled cluster (VCC) method for the calculation of the fundamental vibrational frequencies of polyatomic molecules. The VCC method has shown to give more accurate results for ethylene [99] than the equivalent vibrational configuration interaction (VCI) method [100–102]; a recent review on both VCI and VCC methods is given in [95]. So, it would be interesting to devise a methodology for the construction of PESs that couple the GA with the VCC (or VCI) approach to fit the vibrational spectrum of polyatomic molecules.

Acknowledgments

This work was supported by Fundação para a Ciência e Tecnologia, Portugal, under grant POSC/EIA/55951/2004. JMCM thanks Fundação para a Ciência e Tecnologia (FCT), Portugal, for the sabbatical grant (SFRH/BSAB/636/2006). FVP and CEF thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil for grants. MMA thanks Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, for a grant. We are grateful to the John von Neumann Institut für Computing, Jülich, for the provision of supercomputer time on the IBM Regatta p690+ (Project EPG01).

References

[1] Rydberg R 1931 *Z. Phys.* **73** 376
[2] Klein O 1932 *Z. Phys.* **76** 226
[3] Rees A L G 1947 *Proc. Phys. Soc.* **59** 998
[4] Kosman W M and Hinze J 1975 *J. Mol. Spectrosc.* **56** 93
[5] Vidal C and Scheingraber H 1977 *J. Mol. Spectrosc.* **65** 46
[6] Varandas A J C, Rodrigues S P J and Batista V M O 2006 *Chem. Phys. Lett.* **424** 425
[7] Moré J, Garbow B S and Hillstrom K E 1980 MINPACK: Argonne National Laboratory, 1980 http://www.netlib.org/minpack/ (accessed in October 2007)
[8] Coxon J A and Hajigeorgiou P G 1991 *J. Mol. Spectrosc.* **150** 1
[9] Coxon J A and Hajigeorgiou P G 1992 *Chem. Phys.* **167** 327
[10] Coxon J A and Hajigeorgiou P G 1999 *J. Mol. Spectrosc.* **193** 306
[11] Brühl R, Kapetanakis J and Zimmermann D 1991 *J. Chem. Phys.* **94** 5865

[12] Hedderich H G, Dulick M and Bernath P F 1993 *J. Chem. Phys.* **99** 8363
[13] Grabow J U, Pine A S, Fraser G T, Lovas F J, Suenram R D, Emilsson T, Arunan E and Gutowsky H S 1995 *J. Chem. Phys.* **102** 1181
[14] Seto J Y, Morbi Z, Charron F, Lee S K, Bernath P F and Le Roy R J 1999 *J. Chem. Phys.* **110** 11756
[15] Seto J Y, Le Roy R J, Vergès J and Amiot C 2000 *J. Chem. Phys.* **113** 3067
[16] Huang Y and Le Roy R J 2003 *J. Chem. Phys.* **119** 7398
[17] Le Roy R J 2004 *J. Mol. Spectrosc.* **228** 92
[18] Le Roy R L, Appadoo D R T, Anderson K, Shayesteh A, Gordon I E and Bernath P F 2005 *J. Chem. Phys.* **123** 204304
[19] Coxon J A and Melville T C 2006 *J. Mol. Spectrosc.* **235** 235
[20] Le Roy R L, Huang Y and Jary C 2006 *J. Chem. Phys.* **125** 164310
[21] Bizzocchi L, Guiliano B M, Hess M and Grabow J U 2007 *J. Chem. Phys.* **126** 114305
[22] Li D, Xie F, Li L, Lazoudis A and Lyyra A M 2007 *J. Mol. Spectrosc.* **246** 180
[23] Ogilvie J F 1991 *J. Mol. Spectrosc.* **148** 243
[24] Ogilvie J F 1992 *J. Mol. Spectrosc.* **156** 8
[25] Ogilvie J F 1983 *Comput. Phys. Commun.* **30** 101
[26] Ogilvie J F 1987 *Chem. Phys. Lett.* **140** 506
[27] Ogilvie J F 1994 *J. Phys. B: At. Mol. Opt. Phys.* **27** 47
[28] Ogilvie J F and Liao S C 1994 *Chem. Phys. Lett.* **226** 281
[29] Tiemann E and Ogilvie J F 1994 *J. Mol. Spectrosc.* **165** 377
[30] Ogilvie J F and Molski M 1999 *Spectrochim. Acta* A **55** 1427
[31] Almeida M M, Prudente F V, Maniero A M and Fellows C E 2008 *Chem. Phys. Lett.* submitted
[32] Nakano H 1993 *J. Chem. Phys.* **99** 7983
[33] Nakano H 1993 *Chem. Phys. Lett.* **207** 372
[34] Schmidt M W *et al* 1993 *J. Comput. Chem.* **14** 1347
[35] Gordon M S and Schmidt M W 2005 *Theory and Applications of Computational Chemistry: the First Forty Years* ed C E Dykstra, G Frenking, K S Kim and G E Scuseria (Amsterdam: Elsevier) pp 1167–89
[36] Le Roy R J 2007 LEVEL 8.0: a computer program for solving the radial Schrödinger equation for bound and quasibound levels *University of Waterloo Chemical Physics Research Report* CP-663 http://leroy.uwaterloo.ca/programs/ (accessed in October 2007)
[37] Fellows C E, Verges J and Amiot C 1988 *Mol. Phys.* **63** 1115
[38] Fellows C E 1991 *J. Chem. Phys.* **94** 5855
[39] Piecuch P, Kucharski S A, Kowalski K and Musial M 2002 *Comput. Phys. Commun.* **149** 71
[40] Kendall R A, Dunning T H Jr and Harrison R J 1992 *J. Chem. Phys.* **96** 6796
[41] Wilson A K, van Mourik T and Dunning T H Jr 1996 *J. Mol. Struct. (THEOCHEM)* **388** 339
[42] Tao F M and Pan Y K 1991 *J. Chem. Phys.* **95** 3582
[43] Tao F M and Pan Y K 1991 *J. Chem. Phys.* **95** 9811
[44] Tao F M and Pan Y K 1992 *J. Chem. Phys.* **96** 5815
[45] Halkier A, Helgaker T, Jorgensen P, Klopper W, Koch H, Olsen J and Wilson A K 1998 *Chem. Phys. Lett.* **286** 243
[46] Murdachaew G, Szalewicz K, Jiang H and Bacic Z 2004 *J. Chem. Phys.* **121** 11839
[47] Slavícek P, Kalus R, Paska P, Odvárková I, Hobza P and Malijevský A 2003 *J. Chem. Phys.* **119** 2102
[48] Patkowski K, Murdachaew G, Fou C M and Szalewicz K 2005 *Mol. Phys.* **103** 2031
[49] Herman P R, LaRocque P E and Stoicheff B P 1988 *J. Chem. Phys.* **89** 4535
[50] Hartke B 2004 *Struct. Bond.* **110** 33
[51] Cox G A, Mortimer-Jones T V, Taylor R P and Johnston R L 2004 *Theor. Chem. Acc.* **112** 163
[52] Koskowski F and Hartke B 2005 *J. Comput. Chem.* **26** 1169
[53] Deaven D M and Ho K M 1995 *Phys. Rev. Lett.* **75** 288
[54] Zeiri Y 1995 *Phys. Rev.* E **51** R2769
[55] Niesse J A and Mayne H R 1997 *J. Comput. Chem.* **18** 1233
[56] Hartke B 1999 *J. Comput. Chem.* **20** 1752
[57] Johnston R L 2003 *Dalton Trans.* 4193
[58] Lloyd L D, Johnston R L and Salhi L 2005 *J. Comput. Chem.* **26** 1069
[59] Pereira F B, Marques J M C, Leitão T and Tavares J 2006 *Proc. 2006 IEEE Congress on Evolutionary Computation* vols 1–6 (Vancouver: CEC) pp 2270–7
[60] Pereira F B, Marques J M C, Leitão T and Tavares J 2008 *Advances in Metaheuristics for Hard Optimization, Springer Natural Computing Series* ed P Siarry and Z Michalewicz (Berlin: Springer) pp 223–50
[61] Hennessy M H and Kelley A M 2004 *Phys. Chem. Chem. Phys.* **6** 1085
[62] Meerts W L and Schmitt M 2006 *Int. Rev. Phys. Chem.* **25** 353
[63] Hageman J A, Wehrens R, de Gelder R, Meerts W L and Buydens L M C 2000 *J. Chem. Phys.* **113** 7955
[64] Szydlowska I, Myszkiewicz G and Meerts W L 2002 *Chem. Phys.* **283** 371
[65] Schmitt M, Böhm M, Ratzer C, Vu C, Kalkman I and Meerts W L 2005 *J. Am. Chem. Soc.* **127** 10356
[66] Meerts W L and Schmitt M 2006 *Phys. Scr.* **73** C47
[67] Metzger G J, Patel M and Hu X 1996 *J. Magn. Reson.* B **110** 316
[68] Dods J, Gruner D and Brumer P 1996 *Chem. Phys. Lett.* **261** 612
[69] Ahonen H, de Souza P A Jr and Vijayendra K G 1997 *Nucl. Instrum. Methods Phys.* B **124** 633
[70] Welser L A *et al* 2006 *J. Quant. Spectrosc. Radiat. Transfer* **99** 649
[71] Spalek T, Pietrzyk P and Sojka Z 2005 *J. Chem. Inf. Model.* **45** 18
[72] Makarov D E and Metiu H 1998 *J. Chem. Phys.* **108** 590
[73] Xu Y G and Liu G R 2003 *J. Micromech. Microeng.* **13** 254
[74] Ferreira da Cunha W, Roncaratti L F, Gargano R and Silva G M 2006 *Int. J. Quantum Chem.* **106** 2650
[75] Roncaratti L F, Gargano R and Silva G M 2006 *J. Mol. Struct. (THEOCHEM)* **769** 47
[76] Nocedal J 1980 *Math. Comput.* **35** 773
[77] Liu D and Nocedal J 1989 *Math. Program.* B **45** 503
[78] Deb K and Agrawal R B 1995 Simulated binary crossover for continuous search space *Complex Syst.* **9** 115–48
[79] Deb K and Beyer H G 2001 *Evol. Comput.* **9** 197
[80] Colbert D T and Miller W H 1992 *J. Chem. Phys.* **96** 1982
[81] Prudente F V, Riganelli A and Varandas A J C 2001 *Rev. Mex. Fis.* **47** 568
[82] Korona T, Williams H L, Bukowski R, Jeziorski B and Szalewicz K 1997 *J. Chem. Phys.* **106** 5109
[83] Aziz R A 1993 *J. Chem. Phys.* **99** 4518
[84] Dunning T H Jr 1989 *J. Chem. Phys.* **90** 1007
[85] Woon D E and Dunning T H Jr 1993 *J. Chem. Phys.* **98** 1358
[86] Peterson K A and Dunning T H Jr 2002 *J. Chem. Phys.* **117** 10548
[87] Murrel J N, Carter S, Farantos S C, Huxley P and Varandas A J C 1984 *Molecular Potential Energy Functions* (Chichester: Wiley)
[88] Haykin S 1999 *Neural Networks: A Comprehensive Foundation* 2nd edn (Englewood Cliffs, NJ: Prentice-Hall)
[89] Brown D F R, Gibbs M N and Clary D C 1996 *J. Chem. Phys.* **105** 7597
[90] Prudente F V, Acioli P H and Soares Neto J J 1998 *J. Chem. Phys.* **109** 8801
[91] Bittencourt A C P, Prudente F V and Vianna J D M 2004 *Chem. Phys.* **297** 153
[92] Manzhos S and Carrington T Jr 2006 *J. Chem. Phys.* **125** 084109

[93] Behler J, Lorenz S and Reuter K 2007 *J. Chem. Phys.* **127** 014705
[94] Manzhos S and Carrington T Jr 2007 *J. Chem. Phys.* **127** 014103
[95] Christiansen O 2007 *Phys. Chem. Chem. Phys.* **9** 2942
[96] Toffoli D, Kongsted J and Christiansen O 2007 *J. Chem. Phys.* **127** 204106
[97] Christiansen O 2004 *J. Chem. Phys.* **120** 2140
[98] Christiansen O 2004 *J. Chem. Phys.* **120** 2149
[99] Christiansen O and Luis J M 2005 *Int. J. Quantum Chem.* **104** 667
[100] Bowman J M, Christoffel K M and Tobin F 1979 *J. Phys. Chem.* **83** 905
[101] Christoffel K M and Bowman J M 1982 *Chem. Phys. Lett.* **85** 220
[102] Carter S, Bowman J M and Handy N C 1998 *Theor. Chem. Acc.* **100** 191