Journal of Quantitative Spectroscopy & Radiative Transfer 113 (2012) 929–935

Contents lists available at SciVerse ScienceDirect

Journal of Quantitative Spectroscopy & Radiative Transfer

journal homepage: www.elsevier.com/locate/jqsrt

MARVEL: Measured active rotational–vibrational energy levels. II. Algorithmic improvements ☆

Tibor Furtenbacher, Attila G. Császár *

Laboratory of Molecular Structure and Dynamics, and Dynamics, Institute of Chemistry, Eötvös University, H-1518 Budapest 112, P.O. Box 32, Hungary

ARTICLE INFO

Available online 13 January 2012

Keywords:
MARVEL
Spectroscopic network
Water molecule

ABSTRACT

When determining energy levels from several, in cases many, measured and assigned high-resolution molecular spectra according to the Ritz principle, it is advantageous to investigate the spectra via the concept of spectroscopic networks (SNs). Experimental SNs are finite, weighted, undirected, multiedge, rooted graphs, whereby the vertices are the energy levels, the edges are the transitions, and the weights are provided by transition intensities. A considerable practical problem arises from the fact that SNs can be very large for isotopologues of molecules widely studied; for example, the experimental dataset for the $H_2{}^{16}O$ molecule contains some 160,000 measured transitions and 20,000 energy levels. In order to treat such large SNs and extract the maximum amount of information from them, sophisticated algorithms are needed when inverting the transition data. To achieve numerical effectiveness, we found the following efficient algorithms applicable to very large SNs: reading the input data employs hash codes, building the components of the SN utilizes a recursive depth-first search algorithm, solving the linear least-squares problem is via the conjugate gradient method, and determination of the uncertainties of the energy levels takes advantage of the robust reweighting algorithm.

1. Introduction

There are several fields of science and technology which require the availability of detailed information on the energy level structure of molecules. Computation of accurate thermochemical properties of individual molecules is based on partition functions and the direct summation technique [1–3], as simplified models (e.g., the harmonic oscillator and rigid rotor approximations) are usually not sufficiently accurate. Determination of partition functions, especially at elevated temperatures, requires the availability of a large number of levels impossible to obtain experimentally or can be inaccurate if obtained via simple, perturbative techniques [4–7] of nuclear motion theory. Modelers in combustion [8–10], star-formation [11], planetary atmosphere [12], radiative transfer [13], and chemical vapor deposition [14] research, among others, often need detailed line-by-line information which often only the joint consideration of elaborate spectroscopic measurements [15] and sophisticated quantum chemical computations [16–18] can provide. Recent advances in molecular spectroscopy led to a considerable increase in the extent of experimental high-resolution spectroscopic data, i.e., assigned rovibrational transitions, energy levels, intensities, and line profiles. Some of these data have been deposited, sometimes in a critically evaluated and annotated form, in databases (see, for example, Refs. [19–24]).

☆ This paper is dedicated to three musketeers of French spectroscopy, Professors Jean-Marie Flaud, Claude Camy-Peyret, and Alain Barbe. They, among many other distinguished achievements reflected upon in this volume, suggested a protocol similar to that employed in MARVEL as well as laid some of the ground work for the detailed study of the spectroscopy of water vapor.

* Corresponding author. Tel.: +36 13722929; fax: +36 13722592.
E-mail address: csaszar@chem.elte.hu (A.G. Császár).

0022-4073/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jqsrt.2012.01.005

It is high-resolution molecular spectroscopy which is able to provide the required information experimentally. Experiments are based on high-resolution but narrow-band as well as broad-band but perhaps lower-resolution techniques. The emphasis of the present paper is on the accumulation, validation, handling, visualization, and distribution of the increasing amount of such spectroscopic information on molecular systems [19–28]. Critical evaluation involves computer-aided handling of large-scale databases assembled from all relevant literature sources. Validation of transitions can be based on linelists, preferentially obtained from accurate variational nuclear motion computations. Ultimately, it is the joint utilization of (spectroscopic) experiments, quantum theory, and information technology which helps to secure and extend our firm chemical knowledge on the energy level structure of molecules [16].

Some of the tasks mentioned in the previous paragraph can be helped by recognizing that for individual molecules high-resolution spectra correspond to large-scale deterministic, undirected, (weighted) graphs (networks), made up of energy levels as vertices (nodes), allowed transitions between the levels as edges (links), and weights related to transition intensities. This way one defines what we call spectroscopic networks (SNs) [26,29,30]. The robust organizing principle of SNs is provided by quantum mechanical selection rules; different transitions and transition intensities characterize different spectroscopic techniques. It is important to emphasize that even in the experimentally most thoroughly studied cases the observable transitions form just a tiny part of all the transitions allowed [19,20]. The (sufficiently) complete linelist information about allowed transitions, corresponding to very large SNs, can only be determined via sophisticated variational rotational–vibrational computations [31].

The network-theoretical view of complex SNs, detailed in Ref. [30], offers certain advantages toward completing the characterization of high-resolution molecular spectra. The approximately scale-free property of the overall SN structure [30] leads to the concept of hubs (nodes with a large number of links) and thus straightforwardly to the design of new spectroscopic experiments to improve the information content of SNs. Well-designed experiments help to determine a more accurate and more robust SN with a minimum amount of effort, by preferentially measuring, with improved accuracy, transitions in which less accurately known hubs are involved. Detailed comparison of measured and first-principles hubs helps to determine the "weakest nodes" among the energy levels in an existing experimental SN which, in turn, leads to the identification of transitions which should preferentially be investigated in new experiments designed specifically for their determination. The graph-theoretical view, through the concept of maximum weight spanning trees, should also help to connect components in the measured SN.

Experiments yield relatively small multiedge graphs, with a considerable number of parallel edges, while first-principles computations result in very large simple graphs. The multiedge character of experimental SNs calls for an algorithm which can invert the experimental transition information and yield experimental energy levels ("term values") with well-defined uncertainties. Compared to effective Hamiltonian approaches, still used widely by spectroscopists, involving fitting of a small number of parameters within a complex model to measured spectra, term-value fits involve a large number of parameters, in the form of energy levels, and an exceedingly simple design matrix, based on the Ritz principle. Development of such inversion protocols has a long history in spectroscopy [25,28–42]. It must also be mentioned that inversion techniques, and the resulting weighted least-squares refinements, have helped tremendously other areas of physical chemistry, as well, including thermochemistry [43,44], and reaction kinetics [45,46].

Our own spectroscopic inversion protocol is called MARVEL [29], standing for Measured Active Rotational–Vibrational Energy Levels. The design idea of MARVEL is to separate measurements from models (effective Hamiltonians) and to store the basic experimental observables (rovibronic transitions) and their interdependencies and update the derived data (energy levels) as often as needed. It is important to point out here that the original MARVEL procedure [29] was built upon pioneering work Flaud, Camy-Peyret, and Maillard published quite some time ago [35], and also on the Active Thermochemical Tables (ATcT) approach of Ruscic [43], and a robust reweighting algorithm [47] advocated by Watson [48]. MARVEL is called active since if measurements of new transitions become available they can be added to the network easily and this can be followed by a refinement of the energy levels yielding improved MARVEL energy levels and uncertainties.

Since its invention, the MARVEL technique has been used to handle and validate measured transitions of the following isotopologues of the water molecule: $H_2{}^{17}O$ [19,20], $H_2{}^{18}O$ [19,20], $HD^{16}O$ [20], $HD^{17}O$ [20], and $HD^{18}O$ [20]. Out of the millions of transitions which could technically be detected [49], only relatively few have been observed and assigned experimentally. Of the water isotopologues mentioned, the most extensive set of data is available for $HD^{16}O$ but even in this case only about 55,000 transitions and 9,000 energy levels have been scrutinized by experimentalists. MARVEL has also been used to validate the measured spectra of the parent isotopologue of the ketene molecule [50], characterized by 3194 validated transitions and 1722 validated energy levels. To treat much larger datasets, for example that of $H_2{}^{16}O$ [51], required us to improve all of the algorithms the original MARVEL code employed. Here we report our findings concerning the utility of the algorithms tried and their performance. While the basic facts concerning the MARVEL protocol and the corresponding code, written in the object-oriented C++ language, have been described in Ref. [29], details of the algorithms employed when writing the code have not been provided. It is our hope that this lack of information is remedied here for the benefit of those who would like to develop their own MARVEL-type (active database) procedures.

2. The MARVEL algorithm

The fundamental equation of MARVEL [29] is built upon the Ritz principle, connecting measured wavenumbers (transitions) and energy levels

$$\sigma_{ij} = E_i - E_j, \quad (1)$$

where σ_{ij} is a measured wavenumber ($ij = 1, \ldots, N_t$, where N_t is the number of measured transitions) and E_j is a lower and E_i is an upper rovibronic energy level ($i,j \in 1, \ldots, N_l$, with altogether N_l energy levels taking part in the N_t transitions). Let δ_{ij} be the measurement uncertainty associated with a given transition, σ_{ij}. The principal input to MARVEL is a grand list of N_t experimentally measured, assigned, and labeled transitions with uncertainties and the aim of MARVEL is to determine the N_l energy levels of 'experimental quality' with self-consistent uncertainties.

The principal steps of applying MARVEL to an arbitrary molecule are as follows. (1) Collection of those measured transitions into a database which possess correct assignment, self-consistent labels, and proper uncertainties. (2) Based on the transition data, determination of those energy levels of the given species which belong to a component of a particular spectroscopic network (SN). (3) Within a given SN, set up two vectors containing all the experimentally measured transitions selected and the requested MARVEL energy levels, and an extremely sparse matrix connecting the two vectors, describing the relation between the transitions and the energy levels. The matrix contains for each measured line only two non-zero entries, +1 and −1 for the upper and lower energy levels, respectively. (4) Solution of the set of linear equations corresponding to the chosen pair of vectors and the inversion (design) matrix. During solution of the set of linear equations uncertainties in the measured transitions can be incorporated which results in uncertainties of the energy levels determined. For further characteristics of the MARVEL algorithm the interested reader should consult Ref. [29], henceforth called Part I. In summary, MARVEL is able to yield accurate, reliable, and self-consistent energy levels utilizing all available experimental transition information, correctly reflecting the knowledge present in the input data.

3. Reading and storing the SNs

The measured transitions and the corresponding energy levels in the database can be represented via a graph with perhaps several components. Therefore, we need to store the graph(s), or in other words the spectroscopic network(s) [30], in the most effective way in the computer memory to allow their efficient handling.

There are four widely employed data structures for the representation of graphs in computer memory [52] and [53]: the adjacency list, the incidence list, the adjacency matrix, and the incidence matrix. The last two storage techniques are simple two-dimensional (matrix) representations of graphs, without the possibility to store data characterizing the vertices and edges; therefore, they are not appropriate for the current application. In the adjacency list method vertices are stored as objects and every vertex object contains the list of adjacent vertices. Because this information alone is not sufficient for effective spectroscopic network storage, we employed the incidence list method.

Using the incidence list method we define both vertices and edges as objects. Vertices store their incoming and outgoing edges, and edges store their vertices. While this is certainly not the most memory-efficient algorithm, experimental SNs have much less than a million links and an order of magnitude less nodes causing no storage bottlenecks. This storage technique speeds up the MARVEL procedure so much that it is worth the extra space in memory. In MARVEL, we store the transitions (edges) in a list and the energy levels (vertices) in another list and each transition object contains the upper and lower energy levels as pointers and each energy level object contains a pointer list of those transition objects which connect to the energy level.

The first step during the determination of the energy levels is the reading of the transition database. After reading a line with a format of Eq. (1), a transition object and two (the lower and the upper) energy levels should be created. However, before creating the two energy level objects it needs to be checked whether these energy levels are in the list of the energy level objects or not, as it is necessary to avoid the duplication in the list of energy levels. This step can be very expensive as the size of the energy level list can be more than 10,000. We can speed up this checking if we use a hash code algorithm. This means we create a unique and specific integer number for each energy level and store these integer numbers in a separate list and try to find the given energy level (in fact its hash code) in this integer vector. We used Jenkins' 96 bit-mix function [54] to create the specific hash key from the labels of the energy levels. A general choice for labeling is provided by the usual approximate quantum numbers, the standard normal coordinates (e.g., $\nu_1\nu_2\nu_3$ in the case of water, where ν_1, ν_2, and ν_3 are the stretching, bending, and antisymmetric stretching quantum numbers, respectively) for the vibrational states and the standard rotational notation (e.g., $J_{K_aK_c}$ in the case of an asymmetric top, such as water) for the rotational states. Using the Jenkins' 96 bit-mix function three times for each energy level, we can create only one unique and specific integer number from the six quantum numbers characterizing the energy levels of the water isotopologues.

4. Finding components of the spectroscopic network

The exact value of an energy level within an SN can be determined if and only if there exists a path from this level to the origin (root) of the graph. In spectroscopic networks the root is the lowest-energy level within the given component of the graph (with zero or a well-defined energy value and zero uncertainty). Due to the nature of the (experimental) SNs they may contain transitions or group(s) of transitions which do not connect to the root. Therefore, one has to investigate for each energy level whether it has a connection to the chosen root or not.

In the MARVEL code presented in Part I we used the shortest path algorithm [52,53] to answer this question. In graph theory the shortest path problem is one of the most fundamental network optimization problems; therefore, several algorithms have been developed [52,53]. Three variants of the shortest-path algorithm can be distinguished: the single-destination, the single source (the inverse of the former), and the all-pairs shortest-path problem.

We do not need to find all pairs in the graph when searching a SN; therefore, the all-pairs shortest-path method is certainly not an optimal one. A widely used method for the single-source problem is the Dijkstra algorithm [53]; therefore, we applied this algorithm to build components of the graph in Part I. After extensive testing it turned out that this algorithm is not the best choice for our problem as it results in more information than actually needed and it uses an unnecessary amount of computations for at least two reasons. First, the algorithm checks all energy levels in the database and if the database contains several components, as is often the case, a lot of effort is wasted. Second, the Dijkstra algorithm not only determines but also stores information that is not needed when finding components of the SN. It must be emphasized that if one wants to examine certain properties of the SN, such as its diameter (the average path length in the graph), the Dijkstra algorithm appears to be an excellent choice.

For finding components of the SN, the depth-first search (DFS) algorithm [52,53] appears to be a considerably more efficient algorithm than the Dijkstra algorithm. The DFS algorithm has the very important and useful property that we can build the whole graph from the starting vertex without information on the order of the energy levels included in the transition database. The DFS algorithm is a systematic way to find all the vertices reachable from a selected source vertex. The basic idea of the recursive DFS algorithm can be described as follows. (1) Choose a vertex as a start vertex and mark it as visited. (2) Visit all neighboring vertices, and mark them as visited, using the links of the start vertex. (3) Choose the previously visited vertices as the new starting points. (4) Visit all neighboring vertices of these vertices using their links and mark them as visited. (5) Repeat this until all of the vertices of the graph have been marked as visited.

As to numerical efficiency, one should compare the Dijkstra and the DFS algorithms on the example of a large enough database, like that of $H_2{}^{16}O$ [51] containing close to 160,000 edges. Finding all the components with Dijkstra's algorithm may take several minutes, while that with DFS takes a few thousandth of a second. Therefore, determining the components of the SN is not a bottleneck if one employs the DFS algorithm as now, in contrast to Part I, it requires a trivial amount of CPU time even for the largest SNs envisioned.

5. Solving the system of linear equations

An over-determined system of linear equations

$$\mathbf{aX}=\mathbf{Y} \tag{2}$$

characterizes any given SN, where **Y**, of dimension N_t, and **X**, of dimension N_l-1, contain the measured transitions (σ_{ij}) and the energy levels (E_j), respectively, within the component of the SN considered. The elements of matrix **a**, of dimension $N_t\times(N_l-1)$, are chosen to be −1, +1, or 0 according to the following scheme: −1 or +1 if E_j is the lower or upper level of the *ij*th transition, respectively, and 0 otherwise.

When weights $w_{ij}=\delta_{ij}^{-2}$ are introduced, Eq. (2) becomes

$$\mathbf{AX}=\mathbf{B}, \tag{3}$$

where $\mathbf{A}=\mathbf{a}^T\mathbf{wa}$ and $\mathbf{B}=\mathbf{a}^T\mathbf{wY}$. The dimension of the extremely sparse **A** matrix is $(N_l-1)\times(N_l-1)$ and Eq. (3) is a simple system of linear equations, which can be solved more easily than Eq. (2). If the experimental errors are both random and are characterized by a normal distribution, the solution of Eq. (3) satisfies the maximum likelihood principle. Although **a** is an extremely sparse matrix, computation of **A** and **B** of Eq. (3) is time consuming. Thus, it is worth noting that these matrices can be computed according to the following general analytical formulae:

$$\mathbf{A}=\begin{pmatrix} \sum_k w_{1k} & -w_{12} & -w_{13} & \dots & -w_{1N_t} \\ -w_{12} & \sum_k w_{2k} & -w_{23} & \dots & -w_{2N_t} \\ \dots & \dots & \dots & & \dots \\ -w_{1n_T} & -w_{2n_T} & \dots & & \sum_k w_{N_tk} \end{pmatrix}$$

$$\text{and}\quad \mathbf{B}=\begin{pmatrix} \sum_k w_{1k}\sigma_{1k} \\ \sum_k w_{2k}\sigma_{2k} \\ \vdots \\ \sum_k w_{N_tk}\sigma_{N_tk} \end{pmatrix} \tag{4}$$

The summation indicated for the diagonal elements of **A** means that all uncertainties for transitions in which the first, second, etc. energy level is involved must be summed up. The summation has a similar meaning for **B**. As the **A** matrix is large and extremely sparse (w_{ij} are non-zero only in the few cases where the *i*th and *j*th energy levels are connected by a measured transition), we should store it in a special way taking sparsity into account. The compressed row and column storage (CRS and CCS) formats [55] are the most generally used storage methods for sparse matrices. Therefore, we applied both the CRS and CCS formats to store the **A** matrix. In both methods we store the nonzero matrix elements in three vectors: **val**, **irow**, and **icol**, where the **val** vector contains the values of the nonzero elements. In the CRS method **icol** contains the column indexes of the elements in the **val** vector and irow(*i*) shows the positions in **val** and **icol** of the first element of the *i*th row of the given matrix. The CCS method is very similar to the CRS method but **icol** and **irol** switch roles.

Values of the energy levels can be determined by solving Eq. (3). The resulting vector **X** will contain the energy values. Uncertainties of the energy levels are determined as the appropriate diagonal elements of the inverse of matrix **A**, $\varepsilon_j\approx\sqrt{A_{jj}^{-1}}$, where ε_j denotes the

uncertainty (one standard deviation) associated with the energy level E_j.

Several effective linear solvers are described in the literature. There are two main types of them: direct and iterative [56]. The considerable advantage of the direct methods is that the elements of the inverse matrix can be determined analytically. Their disadvantage is that they are much slower than the iterative algorithms. If we want to determine uncertainties for a large number of energy levels we need a very effective direct solver. Because the **A** matrix is a symmetric positive definite matrix, we can use a sparse-adaptive LDL^T decomposition as a special type of Cholesky linear solvers [56,57]. If approximate uncertainties are sufficient, a possible choice can be the preconditioned conjugate gradient method [58–60], one of the fastest linear solvers. The approximate diagonal elements can be determined using the expression [61]

$$(\mathbf{A}^{-1})_{i,i} = \frac{1}{a_{i,i}} + \frac{1}{a}, \tag{5}$$

where a is the sum of the elements of matrix **A**. The average error of Eq. (5), when compared with the analytical results, is about 5–10%. This error is more than acceptable as generally even a factor of 2 error in the uncertainties is enough to decide whether an energy level is well defined or not. As an illustration, the uncertainty of a very well defined energy level has an uncertainty between $1–10\times10^{-6}$ cm^{-1}, while the uncertainty of an energy level with an average uncertainty is larger than 1000×10^{-6} cm^{-1}.

6. Handling of experimental uncertainties

One of the greatest difficulties when working with a database of the size of a SN is the handling of experimental uncertainties attached to the transitions. The difficulty results from the incorrectness of the uncertainties of some of the measured transitions, which may be due to several factors, such as incorrect assessment of some experimental details, effect of local perturbations, blending within complex experimental features, misassignment, calibration difficulties including misidentified reference lines, etc. Since in a number of cases uncertainties of the measured transitions are too optimistic, the uncertainty values are much less than they should be; therefore, these transitions can distort the values of some of the energy levels (note that understanding error propagation is made extremely difficult by the extremely large number of cycles in SNs). To get realistic energy values we must have realistic uncertainty estimates. Therefore, in cases where the uncertainty is incorrect we must correct it. Thus, we need an algorithm which can detect these problems and correct their uncertainties to the appropriate extent.

The robust reweighting algorithm [47], advocated by Watson [48], seems to be especially well suited for adjusting of the uncertainties of the measured transitions. It is based on the following simple adjustment formula:

$$w_{ij} = \frac{1}{\delta_{ij}^2 + \alpha\,\Delta_{ij}^2} \tag{6}$$

where α is a positive number ($\alpha \le 1/3$) chosen for the given problem and Δ_{ij} is the difference between the original measured and the computed transition

$$\Delta_{ij} = \sigma_{ij} - (E_i - E_j) \tag{7}$$

The choice of α is thoroughly discussed in Ref. [48]. It is sufficient to say here that the smaller α the slower the reweighting becomes. We found in our applications that much smaller values of α than suggested by Watson [48] work fine and thus as the iterative refinement goes along we set α to as small values as 0.01.

A major advantage of the robust reweighting algorithm is that it offers a clear choice when the adjustment of the uncertainties of the lines should be stopped. Adjustment of the uncertainties of the transitions through this iteratively reweighted least-squares scheme can be stopped when the quantity

$$\sum_{ij} \frac{w_{ij}\Delta_{ij}^2}{N_t - N_l} \tag{8}$$

becomes as close to 1 as desired. After applying the robust re-weighting algorithm, MARVEL results in a database containing self-consistent and correctly assigned transitions and the seemingly best possible related uncertainties. Energy levels, and their uncertainties, determined from these transitions are in harmony with the measured transitions and their uncertainties.

Table 1
Characteristics and numerical performance of some of the algorithms employed in the MARVEL code.

Algorithms	Features	Relative CPU times
Reading the input data		
With hash code	No important difference between the two algorithms	1
Without hash code		8
Building the graph		
With DFS	Only graph building	1
With Dijkstra	Lots of extra information	10^4–10^5
Solving linear equations		
With CGM	Numerically exact	1
With LDL^T	Exact	10^2–10^3
Diagonal elements of the $\mathbf{A}^{-1}$ matrix [Eq. (3)]		
With Eq. (5)	Working with 5–10% error	1
With LDL^T	Exact	10^2–10^3

7. Numerical results

During the development of the MARVEL code, as detailed in the previous sections, several different numerical algorithms have been tested. Some of the most relevant ones are given in Table 1.

It is clear from Table 1 that the algorithms perform quite differently when employed to analyze and explore spectroscopic networks. The numerical performance of the methods can be orders of magnitude different, making the identification of the best algorithms straightforward.

Of course, not only the relative performance of the methods is important but also the absolute performance of the best algorithms. In this respect we'd like to point out a few numbers related to the handling of a database containing about 160,000 transition entries and close to 20,000 energy levels. All the numbers below refer to CPU time on a modern laptop using a single processor. Reading the data into memory takes a couple of seconds but this task needs to be performed only once. Once this task is completed, all the other steps in the MARVEL procedure take just a fraction of a second. This means that the efficiency of the present MARVEL procedure is excellent, and this performance allows one to employ MARVEL for the assignment and labeling of any experimental high-resolution molecular spectra with the aid of some computed results. We plan to report such studies in the near future.

8. Conclusions

The number of molecules subjected to MARVEL-type analyses is expected to grow substantially in the near future. These efforts have special relevance as even the most sophisticated quantum chemical techniques are not able to approach the accuracy of experiments for determining the rotational–vibrational energy levels of polyatomic and polyelectronic systems. The same can be said, though to a much lesser extent, about effective Hamiltonian approaches. In favorable cases these can provide data matching the precision of experiments but their predictive power deteriorates considerably when moving to energy regimes not characterized by measured transitions. Clearly, the most viable approach to move toward an information system as complete as possible is to combine incomplete but highly accurate experimental information with complete but inaccurate first-principles results.

The usefulness of MARVEL-type inversion procedures depends heavily on the numerical performance of them. It is reassuring in this respect that the best algorithms applied in this study allow the treatment of indeed very large spectroscopic networks. The largest network treated here contains some 160,000 transitions and 20,000 energy levels and the weighted least-squares determination of the MARVEL energy levels takes just a fraction of a second. To achieve this numerical effectiveness, we found the following algorithms to perform best on SNs. The input data should be read employing hash codes. Building of the components of the SN should employ a recursive depth-first search (DFS) algorithm. Solving the linear least-squares problem becomes especially fast if the conjugate gradient method (CGM) is employed. Determination of the uncertainties of the MARVEL energy levels can be made efficient by the use of a simplified treatment (see Eq. (5)). The robust reweighting algorithm is especially well suited for the adjustment of the uncertainties of the measured transitions.

It appears that the efficiency of the present MARVEL procedure is excellent, and this performance allows one to employ MARVEL for (a) the treatment of the experimental high-resolution spectroscopic data of any molecule of interest, and (b) the assignment and labeling of any experimental high-resolution molecular spectra with the aid of some computed results and some other experimental results.

Acknowledgments

This research was carried out with the financial assistance of the Hungarian Scientific Research Fund (OTKA, K72885 and NK83583). The European Union and the European Social Fund have provided financial support to this project under Grant no. TÁMOP-4.2.1/B-09/1/KMR-2010-0003.

References

[1] Vidler M, Tennyson J. Accurate partition function and thermodynamic data for water. J Chem Phys 2000;113:9766–71.
[2] Wenger Ch, Champion JP, Boudon V. The partition sum of methane at high temperature. J Quant Spectrosc Radiat Transfer 2008;109: 2697–706.
[3] Martin JML, Francois JP, Gijbels R. First principles computation of thermochemical properties beyond the harmonic approximation. I. Method and application to the water molecules and its isotopomers. J Chem Phys 1992;96:7633–45.
[4] Nielsen HH. The vibration–rotation energies of molecules. Rev. Mod. Phys. 1951;23:90.
[5] Mills IM. In: Rao KN, Mathews CW, editors. Molecular Spectroscopy: Modern Research, vol. I. New York: Academic Press; 1972. p. 115.
[6] Allen WD, Yamaguchi Y, Császár AG, Clabo DA, Remington RB, Schaefer HF. A systematic study of molecular vibrational anharmonicity and vibration–rotation interaction by self-consistent-field higher derivative methods. Linear polyatomic molecules. Chem Phys 1990;145:427–66.
[7] Clabo DA, Allen WD, Remington RB, Yamaguchi Y, Schaefer HF. A systematic study of molecular vibrational anharmonicity and vibration–rotation interaction by self-consistent-field higher-derivative methods - asymmetric-top molecules. Chem Phys 1988;123: 187–239.
[8] NATO AGARD (Advisory Group for Aerospace Research and Development). Advisory Report No. 287, Terminology and Assessment Methods of Solid Propellant Rocket Exhaust Signatures, Propulsion and Energetics Panel, Working Group 21 (Brussels: NATO); 1993.
[9] Allen MG. Diode laser absorption sensors for gas dynamics and combustion flows. Meas Sci Technol 1998;9:545–62.
[10] Webber ME, Kim S, Sanders ST, Baer DS, Hanson RK, Ikeda Y. In situ combustion measurements of CO_2 by use of a distributed-feedback diode-laser sensor near 2.0 μm. Appl Opt 2001;40:821–8.
[11] Allard F, Hauschildt PH, Alexander DR, Starrfield S. Model atmospheres of very low mass stars and brown dwarfs. Annu Rev Astron Astrophys 1997;35:137–77.
[12] Pollack JB, Dalton JB, Grinspoon D, Wattson RB, Freedman R, Crisp D, et al. Near-infrared light from Venus' nightside: a spectroscopic analysis. Icarus 1993;103:1–42.
[13] Ramanathan V, Vogelmann AM. Greenhouse effect, atmospheric solar absorption and the earth's radiation budget: from the Arrhenius–Langley era to the 1990 s. Ambio 1997;26:38–46.

[14] Ashfold MHR, May PW, Petherbridge JR, Rosser KN, Smith JA, Mankelevich YA, et al. Unravelling aspects of the gas phase chemistry involved in diamond chemical vapour deposition. Phys Chem Chem Phys 2001;3:3471–85.
[15] Quack M, Merkt F, editors. Handbook of high-resolution spectroscopy. Wiley; 2011.
[16] Császár AG, Fábri C, Szidarovszky T, Furtenbacher T, Mátyus E, Czakó G. The fourth age of quantum chemistry: molecules in motion. Phys Chem Chem Phys 2012;14:1085–106.
[17] Bowman J, Carrington T, Meyer H-D. Quantum approaches for computing vibrational spectra of polyatomic molecules. Mol Phys 2008;106:2145–82.
[18] Tennyson J. Accurate variational calculations for line lists to model the vibration–rotation spectra of hot astrophysical atmospheres. WIREs CMS 2012. doi:10.1002/wcms.94.
[19] Tennyson J, Bernath PF, Brown LR, Campargue A, Carleer MR, Császár AG, et al. IUPAC critical evaluation of the rotational–vibrational spectra of water vapor. Part I. energy levels and transition wavenumbers for $H_2^{17}O$ and $H_2^{18}O$. J Quant Spectrosc Radiat Transfer 2009;110:573–96.
[20] Tennyson J, Bernath PF, Brown LR, Campargue A, Császár AG, Daumont L, et al. IUPAC critical evaluation of the rotational–vibrational spectra of water vapor. Part II. energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$. J Quant Spectrosc Radiat Transfer 2010;111:2160–84.
[21] Jacquinet-Husson N, Scott NA, Chédin A, Crépeau L, Armante R, Capelle V, et al. The GEISA spectroscopic database: current and future archive for Earth and planetary atrmosphere studies. J Quant Spectrosc Radiat Transfer 2008;108:1043–59.
[22] Rothman LS, Gordon IE, Barbe A, Benner DC, Bernath PF, Birk M, et al. The HITRAN 2008 molecular spectroscopic database. J Quant Spectrosc Radiat Transfer 2009;110:533–72.
[23] Müller HSP, Thorwirth S, Roth DA, Winnewisser G. The Cologne database for molecular spectroscopy. CDMS Astron Astrophys 2001;370:L49–52.
[24] Müller HSP, Schröder F, Stutzki J, Winnewisser G. The Cologne database for molecular spectroscopy, CDMS: a useful tool for astronomers and spectroscopists. J Mol Struct 2005;742:215–27.
[25] Lindsay CM, McCall BJ. Comprehensive evaluation and compilation of H_3^+ spectroscopy. J Mol Spectrosc 2001;210:60–83.
[26] Császár AG, Czakó G, Furtenbacher T, Mátyus E. An active database approach to complete rotational–vibrational spectra of small molecules. Annu Rep Comp Chem 2007;3:155–76.
[27] Ruscic B, Boggs JE, Burcat A, Császár AG, Demaison J, Janoshek R, et al. IUPAC critical evaluation of thermochemical properties of selected radicals. Part I. J Phys Chem Ref Data 2005;34:573–656.
[28] Tashkun SA, Perevalov VI, Teffo J-L, Bykov AD, Lavrentieva NN. CDSD-1000, the high-temperature carbon dioxide spectroscopic databank. J Quant Spectrosc Radiat Transfer 2003;82:165–96.
[29] Furtenbacher T, Császár AG, Tennyson J. MARVEL: measured active rotational–vibrational energy levels. J Mol Spectrosc 2007;245:115–25.
[30] Császár AG, Furtenbacher T. Spectroscopic networks. J Mol Spectrosc 2011;266:99–103.
[31] Barber RJ, Tennyson J, Harris GJ, Tolchenov RN. A high-accuracy computed water line list. Mon Not R Astron Soc 2006;368:1087–94.
[32] Aslund N. A numerical method for the simultaneous determination of term values and molecular constants. J Mol Spectrosc 1974;50:424–34.
[33] Albritton DL, Harrop WJ, Schmeltekopf AL, Zare RN, Crow EL. A critique of the term value approach to determining molecular constants from the spectra of diatomic molecules. J Mol Spectrosc 1973;46:67–88.
[34] Plíva J, Telfair WB. Correlations and accuracy of estimation of spectroscopic constants and term values. J Mol Spectrosc 1974;53:221–45.
[35] Bykov AD, Naumenko OV, Pshenichnikov AM, Sinitsa LN. Shcherbakov AP. An expert system for identification of lines in vibrational–rotational spectra. Opt Spectrosc 2003;94:528–37.
[36] Flaud J-M, Camy-Peyret C, Maillard JP. Higher ro-vibrational levels of H_2O deduced from high resolution oxygen–hydrogen flame spectra between 2800–6200 cm^{-1}. Mol Phys 1976;32:499–521.
[37] Moruzzi G, Strumia F, Lees RM, Mukhopadhyay I. CH_3OH laser lines assignments and frequency prediction. Infrared Phys 1991;32:333–47.
[38] Moruzzi G, Xu L-H, Lees RH, Winnewisser BP, Winnewisser M. Investigation of the ground vibrational state of CD_3OH by a new "Ritz" program for direct energy level fittting. J Mol Spectrosc 1994;167:156–75.
[39] Lavrov BP, Umrikhin IS. Optimal values of rovibronic energy levels for triple electronic states of molecular deuterium. J Phys B: At Mol Opt Phys 2008;41:105103.
[40] Öberg K. Isotope shifts and accurate wavelengths in Ne II and Ne III. Eur Phys J D 2007;41:25–47.
[41] Kramida AE. The program LOPT for least-squares optimization of energy levels. Comput Phys Commun 2011;182:419–34.
[42] Watson JKG. The use of term-value fits in testing spectroscopic assignments. J Mol Spectrosc 1994;165:283–90.
[43] Ruscic B, Pinzon RE, Morton ML, Von Laszevski G, Bittner SJ, Nijsure SG, et al. Introduction to active thermochemical tables: several "key" enthalpies of formation revisited. J Phys Chem A 2004;108:9979–97.
[44] Császár AG, Furtenbacher T. From a network of computed reaction enthalpies to atom-based thermochemistry (NEAT). Chem Eur J 2010;16:4826–35.
[45] Feeley R, Seiler P, Packard A, Frenklach M. Consistency of a reaction dataset. J Phys Chem A 2004;108:9573–83.
[46] Frenklach M, Packard A, Seiler P, Freeley R. Collaborative data processing in developing predictive models of complex reaction systems. Int J Chem Kinet 2004;36:57–66.
[47] Huber PJ. In: Robust statistics. New York: Wiley; 1981.
[48] Watson JKG. Robust weighting in least-squares fits. J Mol Spectrosc 2003;219:326–8.
[49] Furtenbacher T, Császár AG. The role of intensities in determining characteristics of spectroscopic networks. J Mol Struct, doi:10.1016/j.molstruc.2011.10.057, in press..
[50] Fábri C, Mátyus E, Furtenbacher T, Nemes L, Mihály B, Zoltáni T, et al. Variational quantum mechanical and active database approaches to the rotational–vibrational spectroscopy of ketene. J Chem Phys 2011;135:094307.
[51] Tennyson J, Bernath PF, Brown LR, Campargue A, Császár AG, Daumont L, et al. IUPAC critical evaluation of the rotational-vibrational spectra of water vapor. Part III. energy levels and transition wavenumbers for $H_2^{16}O$. J Quant Spectrosc Radiat Transfer 2012 in preparation.
[52] Sedgewick R, editor. Algorithms in C++. Part 5: Graph algorithms. Addison-Wesley; 2002.
[53] Diestel R. Graph Theory. 3rd ed. Berlin: Springer; 2005.
[54] Jenkins RA. Hash function for hash table lookup. ⟨http://burtleburtle.net/bob/hash/doobs.html⟩.
[55] Bai JDZ, Dongarra J, Ruhe A, Vorst H. Templates for the solution of algebraic eigenvalue problems: a practical guide. Soc Ind Appl Math 2000.
[56] Dahlquist G, Björck Å. Numerical Methods. Dover 2003.
[57] Davis TA. Direct methods for sparse linear systems. SIAM; 2006.
[58] Golub GH, Loan CF. Matrix computations. The Johns Hopkins University Press; 1996.
[59] Paige CC, Saunders MA. Algorithm 583. LSQR: sparse linear equations and least squares problems. ACM Trans Math Softw 1982;8:43–71.
[60] Paige CC, Saunders MA. LSQR: An algorithm for sparse linear equations and sparse least squares. ACM Trans Math Softw 1982;8:195–209.
[61] Simons G, Yao Y. Approximating the inverse of a symmetric positive definite matrix. Linear Algebra Appl 1998;281:97–103.