SEARCHING IN UV/VIS SPECTRAL LIBRARY

D. Hristozov, P. Penchev*, G. Andreev

Department of Analytical Chemistry, University of Plovdiv, 4000 Plovdiv, Bulgaria

Abstract

The program system *UVLib* for searching in UV/Vis spectral databases is described. It is a user friendly menu driven program working in Microsoft Windows environment. Four different search algorithms are implemented in the program: the sum of the squared absorbance differences, sum of the absolute absorbance differences, scalar product of two spectral vectors, and correlation coefficient between two spectral arrays. Their implementation is consistent with peculiarities of the computer representation of UV/Vis spectra.

Introduction

The development of automated systems for structure elucidation and identification of organic compounds continues to attract the attention of spectroscopists. The spectroscopic techniques used for this purpose are ^{1}H and ^{13}C NMR spectroscopy, low resolution mass spectrometry and infrared spectroscopy [1]. The UV/Vis spectra are mainly used for quantitative determination of organic compounds in their solutions [2]. Typically the absorption spectra in 200 - 900 nm wavelength interval have one to three spectral bands due to electron transitions in the molecule. These bands are wide enough to obscure the "characteristic" absorbance of chemical substructures but in spite of that, the proper identification of the compound under study can be obtained if the searched library contains the corresponding compound's spectrum [3].

In this paper we describe several spectral similarity measures implemented in a Windows based program. Their implementation is consistent with peculiarities of the computer representation of UV/Vis spectra: different spectral intervals, concentration's influence on absorbance values, solvent's shift of spectrum maxima, and so on.

System description

The *UVLib* is a user friendly program working under Windows 95 (or higher) environment. The program code is written in Object Pascal and uses the Delphi 5 support of data bases. The main program window is shown in figure1.

The program can maintain unlimited number of spectral libraries and individual compound spectra. Each entry in a library is represented by the following data: chemical name, Wiswesser line-formula chemical notation, molecular formula, nominal molecular mass, chemical structure, solvent description and spectral data. The latter are full spectral curves of absorption values at 1 nm with spectral regions varying in the 180 - 900 nm range. The absorbance values are stored as four byte integers representing the absorbance in the range 0.0 - 0.999 a.u. A middle-size working data base is composed consisting of 1086 spectra [4].

* Corresponding author, e-mail address: plamen@argon.acad.bg

The implementation of a so called *flat-file* data base gives fast access to a library entry through sorting them by a user-selected entry's field as well as by a sophisticated query concerning one or more fields of the library entries.

The unknown UV/Vis spectrum can be loaded from a simple text file containing a sample description and pairs of wavelength and absorbance values.

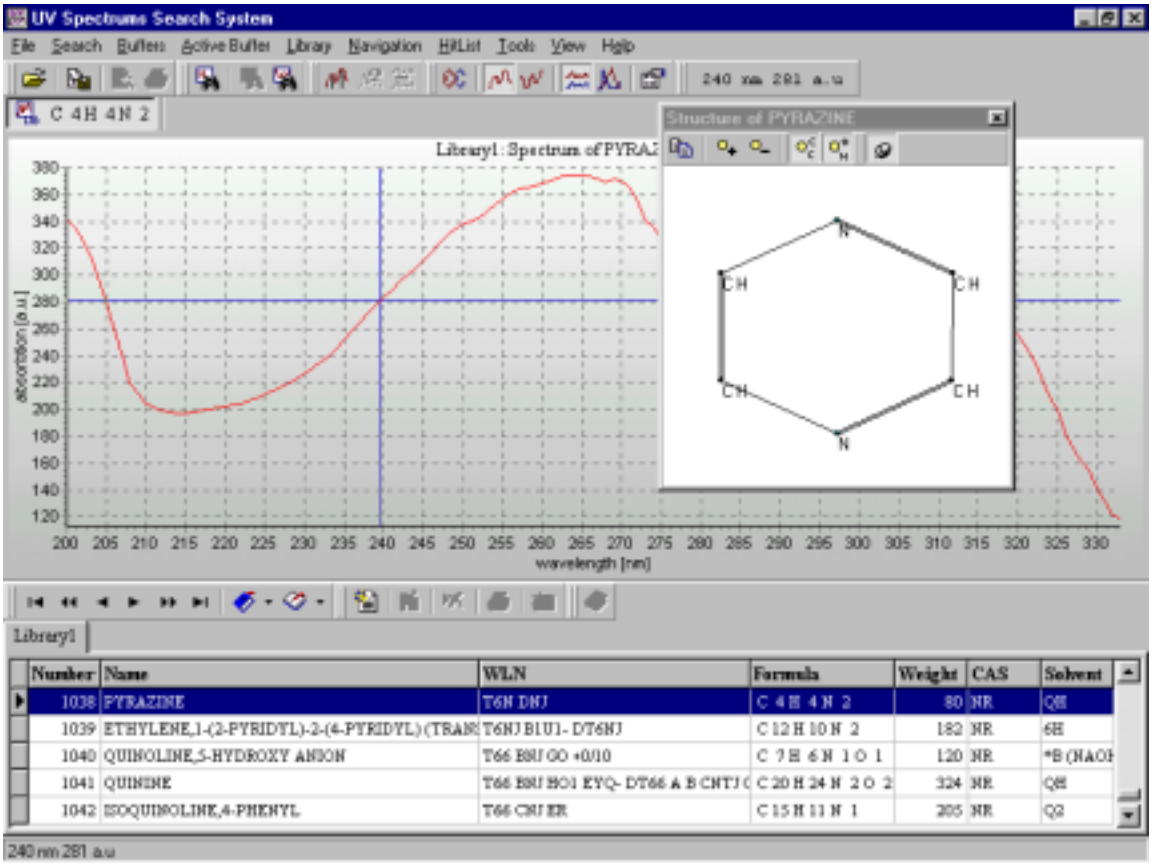

Figure 1. The main program window.

Search algorithms

Four different measures (*hit quality indices* HQI_1 to HQI_4) have been used to describe the similarity between UV/Vis spectra. All four HQIs range between zero and 999: the last value is obtained for identical spectra. Let N be the number of absorbance values used for curve matching and A_k^U and A_k^R be the absorbances at the k^{th} wavelength in the spectrum of the unknown and in that of the reference (library) spectrum, respectively.

Hit quality index HQI_1 is based on the *sum of the squared absorbance differences*, S_1, equation (1).

$$HQI_1 = 999\,(1 - S_1) \qquad \text{with} \quad S_1 = \sqrt{\sum_k (A_k^U - A_k^R)^2 / N} \qquad (1)$$

Hit quality index HQI_2 is calculated from *the sum of the absolute absorbance differences*, S_2, equation (2).

$$HQI_2 = 999\,(1 - S_2) \qquad \text{with} \quad S_2 = (1/N)\sum_k \left|A_k^U - A_k^R\right| \qquad (2)$$

Hit quality index HQI_3 is the *scalar product of two spectral vectors*, equation (3).

$$HQI_3 = 999\, S_3 \qquad \text{with} \quad S_3 = \frac{\sum_k A_k^U A_k^R}{|A^U| \cdot |A^R|} \tag{3}$$

Hit quality index HQI_4 is based on the *correlation coefficient between spectral arrays*, equation (4).

$$HQI_4 = 999\,(S_4 + 1)/2 \qquad \text{with } S_4 = \frac{\sum_k (A_k^U - \overline{A^U})(A_k^R - \overline{A^R})}{\sqrt{\sum_k (A_k^U - \overline{A^U})^2 * \sum_k (A_k^R - \overline{A^R})^2}} \tag{4}$$

Straightforward application of these four measures with whole spectra is impossible because of the variable wavelength regions of library spectra. Also the query spectrum is usually measured and stored with the wavelength region different from that of the corresponding reference spectrum. That is why the user is presented with a dialogue window for setting up the spectral interval for curve matching. The default matching interval is that of the query spectrum, and the user can select a subregion of it. Our experience shows that the best subregion is usually the original one narrowed by 10 nm to 20 nm at both ends. Of course, the user can use his/her own experience to select the matching interval, e.g. by excluding a (less informative) flat region in the spectrum.

The other difficulty with first two measures is that they depend on the absolute values of absorbance. The representation of UV/Vis spectra adopted in hard-copy collections is as a raw absorbance curves which reflect sample concentration. We also preserved the original spectral curves as the sample concentration is given in nearly half of the entries. This led to an impossibility of straightaway application of equations (1) and (2). To make these two comparisons reliable the user can select the preliminary normalisation (A_k/A_{max}) of both spectral curves to be done before the application of equations (1) and (2). This normalisation has not a mathematical effect on equations (3) and (4), and is therefore not applied with them. Another reliable option is to remove the impact of the background line by shifting the spectra by A_{min} value ($A_k - A_{min}$). The user can select the first, second or both options: the last gives the so called scale normalisation $(A_k - A_{min})/(A_{max} - A_{min})$.

Acknowledgement

We are grateful to the late J.T. Clerc (ETH Zurich, Switzerland) for making this database available to us in an appropriate text files.

References

1. N.A.B. Gray; *Computer-Assisted Structure Elucidation.* John Wiley, New York, 1986.
2. H.-H. Perkampus; *UV-VIS Spectroscopy and Its Application*. Springer-Verlag, Berlin, 1992.
3. C.W. Brown, S.M. Donahue; *Searching a UV-Visible Spectral Library*. Appl. Spectrosc., 42, 347-352, 1988.
4. The spectra are part of private spectra collection of the late Prof. J.T. Clerc, ETH Zurich, Switzerland.