Infrared spectra interpretation by means of computer

Plamen N. Penchev*, Nikolay T. Kotchev, George N. Andreev

Department of Analytical Chemistry, Chemical Faculty, University of Plovdiv, BG-4000 Plovdiv, Bulgaria

Abstract

A computer system for interpretation of infrared spectra is described. The methods implemented in the systems are: search in spectral libraries, step-by-step analysis of mixture spectra, and classification of infrared spectra with the aid of linear discriminant analysis, artificial neural networks, and k-nearest neighbors. To illustrate work of the system an example of methods application is included.
Keywords: Infrared spectra, Chemometrics, Artificial neural networks, Library search.

INTRODUCTION

Structure elucidation of chemical compounds from their spectra is one of the main tasks of infrared spectroscopy. For this purpose a number of approaches has been developed: expert systems, library search in spectroscopic data bases, and various pattern recognition techniques [1]. All these approaches suffer from drawbacks and limitations that could be overcome only by a combined application of several chemometric approaches [2].

The purpose of this paper is to describe the joint application of several methods for computer-assisted interpretation of infrared spectra. All methods - seven different library search routines, step-by-step regression analysis of hitlist spectra, and classification of IR spectra with the aid of expert-knowledge, linear discriminant analysis (LDA) and artificial neural networks (ANNs) - are implemented in the infrared spectra search system *IRSS* [3-5]. An example of methods' application is given to illustrate the spectra interpretation process.

METHODS

System Description. A depiction of the system as it currently exists is given in Figure 1. The main task of the program system *IRSS* is compound identification through library search routines. The user can apply one of the seven implemented search algorithms to match spectra of unknown compounds against the library spectra [2-4]. For this purpose two spectral libraries were created: *PU-Lib* (Plovdiv Uni Library*)* consisting of 608 spectra measured in our laboratory [3], and *CC1000* with 1000 spectra purchased from Chemical Concepts [6]. Databases are composed of IR spectra, structural data, molecular formulas, and compound names. The spectral range is from 500 cm^{-1} to 3700 cm^{-1} with a sampling interval of 4 cm^{-1}, corresponding to 801 data points; the latter are absorbance values represented as 8

*Correspondence author, e-mail: plamen@argon.acad.bg, fax: +359-32-635-049.

bit numbers. The LDA, ANN and KNN classifiers were prepared by means of our software in the Laboratory for Chemometrics at Technical University of Vienna with 13484 spectra (*CC13484* library) delivered by Chemical Concepts [6].

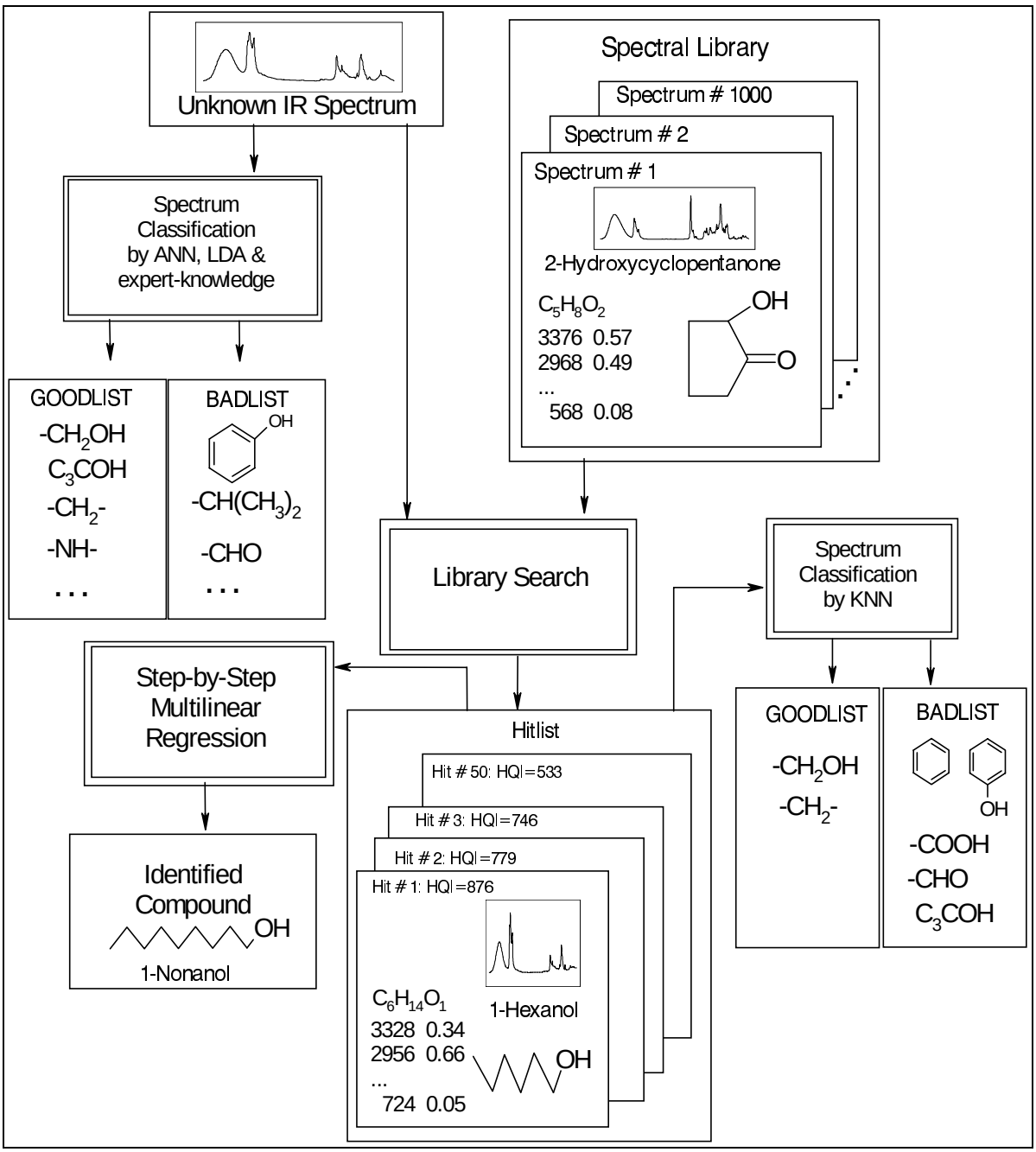

Figure 1. Scheme of the application of all implemented methods in *IRSS*. Single-line boxes denote data and obtained results, others denote algorithms.

When library search fails to identify the compound the user can proceed in two ways [2]: (1) he/she can apply step-by-step regression between the hitlist spectra and studied spectrum in order to obtain reliable compound identification if its spectrum is among the first several hits, or (2) he/she can classify the IR spectrum in order to derive a list of probable substructures which are present or absent in the compound under study. For this purpose more than 70 spectral classifiers have been created

that are based on expert knowledge, linear discriminant analysis and artificial neural networks [5]. These three types of spectral classifiers are implemented in the program module *IRIS* which uses directly peak table of the spectrum in the active buffer of the program *IRSS*.

Another possibility for structure elucidation is the analysis of the structures of the hitlist obtained by a library search. As the hitlist contains structural information relevant to the unknown structure the user can apply a set of classifiers that use the k-nearest neighbor (KNN) method [7]. The KNN classification is performed by a module implemented directly in the program IRSS.

The results of the classification according to substructures represent so called GOODLIST (present substructures) and BADLIST (absent substructures) [8]. The substructures can be fed as input into an isomer generator software which generates a set of all plausible compound structures. For this purpose one can use the program *MOLGEN* [9] which computes complete and redundancy free sets of connectivity isomers for a given brutto formula.

Library search system *IRSS*. This is a Windows based program for performing library search in infrared spectral data bases; the implemented methods in its old versions are described previously [2-4]. The main features of the system are: peak search using three algorithms - forward, reverse, and scalar product; full-curve spectrum search applying four algorithms - sum of squared differences, sum of absolute value differences, scalar product, and correlation coefficient; interactive analysis of spectra of mixtures with the aid of step-by-step multilinear regression with increasing number of hitlist entries; creating user-generated libraries, as well as deleting and merging of libraries, and adding or removing spectra to/from a library.

Binary mixture analysis. This is a procedure for qualitative analysis of spectra of binary organic mixtures [2]. It is based on the calculation of so called pseudo-concentrations of the components in a mixture (matrix $C_{1,H}$); all subscripts of matrices express their dimensions. The pseudo-concentrations are calculated from the spectra in the hit list ($S_{N,H}$) and mixture spectrum ($M_{1,N}$) according to the equation:

$$C_{1,H} = M_{1,N}\, S_{N,H}^{T}\, (S_{H,N}\, S_{N,H}^{T})^{-1}; \qquad (1)$$

where the superscripts "*T*" and "*-1*" denote a transposed and inverse matrix, and the subscripts *N* and *H* are the number of spectral points and the number of hitlist spectra involved in the calculations, respectively. The calculations are performed with increasing *H*, and the program presents graphs $C_K = f_K(H)$ where *K* designates the number of the corresponding hitlist compound. The user can decide which compounds are components of the mixture by comparing the relative stability of the corresponding curves, or by using some statistical criteria [2] (see also the example of application).

***IRIS*.** This is a program module for infrared spectra classification. It operates under Windows environment and can be started directly from the library search system *IRSS*. It uses chemometric methods and expert-knowledge for classification of IR spectra according to the presence or absence of a set of chemical substructures. Both implemented chemometric methods, LDA and ANN, were described in detail

in [2,5]. The expert-knowledge classification is based on the concept of characteristic intervals. The latter were derived from literature sources, e.g. [10] and the citations therein, and were corrected through our experience [11].

KNN classification. It evaluates library search results. The hitlist obtained by a spectral search is processed to determine the number of occurrences of each substructure to which the spectrum is classified. Further, an algorithm is applied to determine conditional probabilities for the substructure presence or absence. It is based on the theory of binary classification initially applied by us to IR spectra with LDA and ANNs [5]; the only difference here is that the ANN (or LDA) output value is replaced by the number of hits containing the corresponding substructure.

Programming, hardware requirements, and availability. The program codes of *IRSS* and *IRIS* were written and compiled in Borland Pascal. The programs were tested on 100% IBM compatible PC computers. The operating environments Windows 3.1x and Windows 95 were fully tested. University scientists and users from non-profit organizations can get a free copy of the executive files and *PU-Uni* library via Internet at the address: http://www.argon.acad.bg/plamen/IRSS.html.

EXAMPLE OF APPLICATION

As an example we will consider classification of the IR spectrum of 1-nonanol from PU-Lib library. The results of the application of all developed methods are schematically represented in Figure 1. As could be seen the applied search method - forward peak search - fails to identify the compound: it appeared on the second hitlist position. The user can apply step-by-step regression which is developed to identify the components of a given mixture (here only one component). Its application to the first 10 spectra from the forward-peak-search hitlist gives the results presented in Figure 2. The curve of the second hit is the most stable one with r.s.d. = 9.8%; for comparison, the curve of the first hit (1-hexanol) gives r.s.d. = 332.6%.

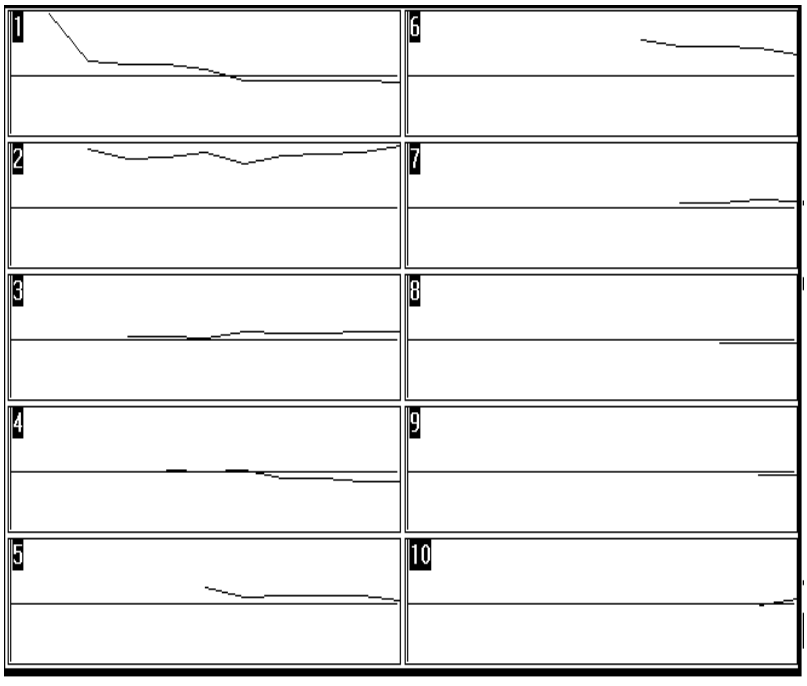

Figure 2. Graphical representation of mixture analysis results as they are given by the program IRSS: graphs $C_K = f_K(H)$ for the first ten hits. 1-nonanol is searched in CC1000 library with forward-peak-search method and appeared as a second hit. The straight lines in the middle of the windows are the values with zero pseudo-concentrations.

The results obtained by the classification of query spectrum with the aid of LDA and ANN classifiers are presented in Table 1a. All these results are correct but those with the aid of expert-knowledge show some errors: hyper-predictions of secondary amine/amide, secondary and tertiary alcohols and isopropyl group. Here also, all negative predictions are correct.

In Table 1b are given the results from application of a set of 20 KNN classifiers on the hitlist obtained by full-curve spectral search with correlation-coefficient similarity measure in CC1000 library. As can be seen all predictions are correct and the primary alcohol fragment is predicted.

Table 1. The results (threshold precision 95%) from the classification with the aid of a) LDA and ANN classifiers, and b) KNN ones. For the latter classification 50 hits are processed from the hitlist obtained by correlation-coefficient full-curve search in CC1000 library.

a)

IR spectrum of 1-Nonanol as 'unknown'			
Classifier	Y/N	Prec	C o m m e n t s
-CH2-OH	Y	97	primary alcohol, ANN
>CH2	Y	99	methylene group, LDA
>CH2	Y	97	methylene group, ANN
-CH=CH2	N	95	vinyl, LDA
-CH=CH2	N	97	vinyl, ANN
-C(CH3)3	N	97	tertiary butyl, LDA
-C(CH3)3	N	97	tertiary butyl, ANN
C2CH-OH	N	98	secondary alcohol, ANN
phenol	N	99	phenol, arbitrary substituted; LDA
phenol	N	96	phenol, arbitrary substituted; ANN
phenyl	N	99	mono substituted benzene, LDA
phenyl	N	99	mono substituted benzene, ANN
-COOH	N	95	carboxylic acid, ANN
benzene	N	99	benzene, arbitrary substituted, LDA
benzene	N	98	benzene, arbitrary substituted, ANN
p-benzene	N	98	para substituted benzene, LDA
m-benzene	N	99	meta substituted benzene, LDA
m-benzene	N	95	meta substituted benzene, ANN
o-benzene	N	96	ortho substituted benzene, LDA
o-benzene	N	95	ortho substituted benzene, ANN
-CHO	N	95	aldehyde, LDA

b)

IR spectrum of 1-Nonanol as 'unknown'			
Classifier	Y/N	Prec	C o m m e n t s
>CH2	Y	99	methylene, KNN
-CH2-OH	Y	96	primary alcohol, KNN
o-benzene	N	96	ortho-benzene, KNN
acid	N	95	carboxylic acid, KNN
C3COH	N	99	tertiary alcohol, KNN
phenol	N	99	phenol, KNN

ACKNOWLEDGEMENTS

P.N.P. thanks Prof. Kurt Varmuza from Technical University of Vienna for giving him an access to the SpecInfo IR database, as well as for many stimulating discussions.

REFERENCES

1. H.J. Luinge. Automated Interpretation of Vibrational Spectra. Vib. Spectrosc., **1** (1990) 3.

2. P.N. Penchev. Ph. D. Thesis. Plovdiv, Bulgaria, **1998** (there is a copy of the dissertation as hypertext document in Bulgarian language at http://argon.acad.bg/plamen/Thesis/Contents.htm).

3. P.N. Penchev; A.N. Sohou; G.N. Andreev. Description and Performance Analysis of an Infrared Library Search System. *Spectrosc. Lett.,* **29** (1996) 1513.

4. P.N. Penchev; N.T. Kotchev; G.N. Andreev. IRSS: A Program System for Infrared Library Search. *Comp. Rend. Acad. Bulg. Sci*., **51** (1998) 67.

5. P.N. Penchev; G.N. Andreev; K. Varmuza. Automatic Classification of Infrared Spectra Using a Set of Improved Expert-based Features. *Anal. Chim. Acta*, **388** (1999) 145.

6. Chemical Concepts, P.O. Box 10 02 02, D-69442 Weinheim, Germany.

7. P.N. Penchev ; N.T. Kotchev. Unpublished results, **1998**.

8. B.G. Buchanan; G.L. Sutherland; E.A. Feigenbaum. Heuristic DENDRAL: a Program for Generating Explanatory Hypotheses in Organic Chemistry, pp. 209-254, in: Meltzer, B. and Michie, D. (Eds.), *Machine Intelligence **4***, Edinburgh University Press, Edinburgh, **1969**.

9. C. Benecke; R. Grund; R. Hohberger; A. Kerber; R. Laue; T. Wieland. MOLGEN$^+$, a Generator of Connectivity Isomers and Stereoisomers for Molecular Structure Elucidation. *Anal. Chim. Acta*, **314** (1995) 141.

10. E. Pretsch; T.J. Clerc; J. Seibl; W. Simon. *Tables of Spectral Data for Structure Determination of Organic Compounds*. Springer, Berlin, **1989**.

11. G.N. Andreev; O.K. Argirov; P.N. Penchev. Expert system for the interpretation of infrared spectra. *Anal. Chim. Acta*, **284** (1993) 131.