*Asian Chemistry Letters* Vol 13, Nos 3 & 4 (2009) 195-200

# Computer-assisted structure elucidation of organic compounds by infrared spectroscopy*

P N Penchev[a], G N Andreev[a] and K Varmuza[b]

[a]*Faculty of Chemistry, University of Plovdiv, 24 Tsar Assen Str, BG-4000 Plovdiv, Bulgaria*

[b]*Vienna University of Technology, Laboratory for Chemometrics, Getreidemarkt 9/166, A-1060 Vienna, Austria*

***This article is dedicated to Professor Hiroaki Takahashi, Waseda University – Tokyo, on the occasion of his 75th anniversary**

The joint application of several chemometric methods for infrared spectra interpretation/classification is described. The methods are implemented in several programs working under MS-DOS and Windows environment. They include: 1) seven routines for searching in spectral libraries; 2) binary classification with artificial neural networks, linear discriminant functions and expert-knowledge classifiers; 3) processing of library search results (a hitlist) with the concept of maximum common substructure and with a set of k-nearest neighbor classifiers. As an illustration of the structure elucidation process an example is given.

## 1 Introduction

Structure elucidation of chemical compounds from their spectra is one of the main tasks of infrared spectroscopy. Automation of this activity by means of computers enhances considerably the performance and reliability of spectra interpretation. For this purpose a great number of approaches have been developed including: expert systems, library search in spectroscopic data bases, and various pattern recognition techniques [1-3]. Some of these methods classify the infrared spectra according to the presence or absence of a set of preliminarily defined substructures or more general structural properties in a molecule. Several examples are the prediction of 3D molecular structures from IR data [4], the use of fuzzy logic [5], the automatic generation of a knowledge base from IR data [6], investigations of IR-spectrum-structure correlations by Kohonen and counterpropagation neural networks [7] and multiple layered perceptron networks [8].

All approaches mentioned above attempt to resolve the abstractly formulated relationship:

$$spectrum = f(structure) \tag{1}$$

Besides the principal impossibility mentioned by Clerc [9] to find the inverse function $f^{-1}$ also other drawbacks common to all computer-based approaches exist. The most important of them are: (1) the infrared spectrum does not reflect all aspects of the compound's electronic structure and nuclear configuration; (2) the spectral features used by the multivariate classification (or interpretation) restrict the structure information embedded in the IR spectrum; (3) the learning set used by training, or the searched spectral libraries are not representative enough; (4) there exist a number of limitations of the model applied for classification. In our opinion the limitation (2) to (4) can be overcome only by a combined application of the approaches mentioned above; the first drawback is surmounted only by a joint use of several spectroscopic techniques [1,2].

In this paper we describe and discuss a combined application of several methods for computer-assisted structure elucidation of organic compounds based on IR spectra: library search routines, maximum common substructure (MCS) analysis of hitlist structures, and classification of IR spectra with the aid of expert knowledge, linear discriminant analysis (LDA) and artificial neural networks (ANNs). All methods

*Corresponding author :*
*e-mail: plamen@uni-plovdiv.bg* (Dr P N Penchev)

except MCS analysis are implemented in the infrared spectra search system *IRSS* [10,11]. The MCS analysis is performed with the external program ToSiM [12,13], but can be applied to the structures of a hitlist obtained by the library search with *IRSS*. Also, an example is given to illustrate the spectra interpretation process.

## 2 Methods

### *2.1 System Description*

A depiction of the system as it currently exists is given in Fig. 1. The main task of the program system *IRSS* is compound identification through library search routines. The user can apply one of seven implemented search algorithms to match the spectrum of an unknown compound against the library spectra [9, 10]. For this purpose four spectral libraries were created for use by the system: *IRLIB01*, *IRLIB02* and *IRLIB03* consisting of 618 spectra measured in our laboratory [10], and *CC13484* with 13484 spectra provided

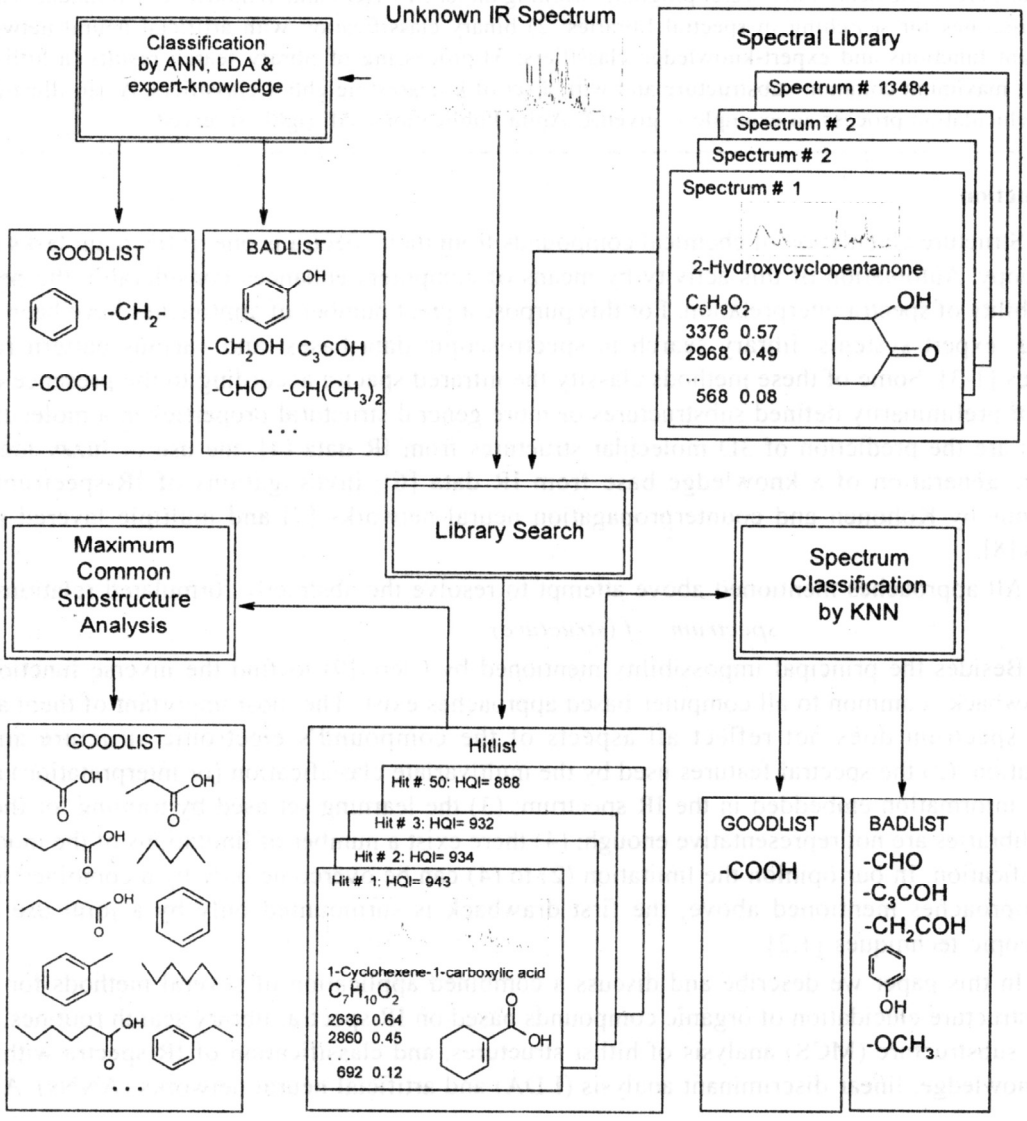


Fig. 1. Scheme of the methods implemented in IRSS. Single-line boxes denote data and obtained results, others denote algorithms. Results shown have been obtained for the example using an IR spectrum of 4-phenyl-butanoic acid as an unknown

by Chemical Concepts [14]. The databases are composed of IR spectra, structural data, molecular formulas, and compound names. The spectral range is from 500 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 bit numbers.

When library search fails to identify a compound the user 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 which are based on expert knowledge, linear discriminant analysis and artificial neural networks [15]. These three types of spectral classifiers are implemented in the program module *IRIS* which uses directly the peak table of a spectrum.

Another approach for structure elucidation using infrared spectra is an analysis of the structures in the hitlist obtained by library search. As the obtained hitlist contains structural information relevant to the unknown structure [9] the user can apply the concept of maximum common substructure [13,16] or a set of classifiers which use the k-nearest neighbor (KNN) method. The KNN classification is performed by the module *IRIS*.

From the results of an MCS analysis and of spectra classifications a so called GOODLIST (present substructures) and a BADLIST (absent substructures) can be built [17]. The substructures can be fed as input into an isomer generator software which generates a set of all possible compound structures. An appropriate isomer generator is the program *MOLGEN* [18] running under MS-Windows which computes complete and redundancy free sets of connectivity isomers for a given brutto formula. The applicable structural restrictions in *MOLGEN* relevant to this work are: GOODLIST (overlapping or not overlapping substructures), BADLIST, lower and upper limits for bond multiplicity and ring size.

### *2.2 Library Search System IRSS*

This is a Windows based program for performing spectral similarity searches in infrared spectral data bases. The main features of the system are [11,19]:

- peak search using three algorithms: forward, reverse, and scalar product of peaks;
- full-curve spectrum search applying four different algorithms: sum of squared differences, sum of absolute value differences, scalar product, and correlation coefficient;
- search for compound name;
- interactive analysis of spectra of mixtures with the aid of a step-by-step multilinear regression with automatically increasing the number of hitlist entries (the latter are regarded as possible mixture components);
- creating user-generated libraries, as well as deleting and merging of libraries, and adding or removing spectra to/from a library.

### *2.3 Infrared Spectra Interpretation System (IRIS)*

This is a program module for infrared spectra interpretation. It operates under Windows environment and can be started directly from the library search system *IRSS*. It uses chemometric methods and expert-knowledge for the classification of IR spectra with the aim to recognize presence or absence of a set of chemical substructures. Both implemented chemometrics methods, LDA and ANN, have been described in detail elsewhere [15,19]. The expert-knowledge classification is based on the concepts of characteristic intervals. The latter are derived from literature, e.g. [20] and citations therein, as well as from our experience [21]. The conclusion for a given substructure is obtained as a result of 'AND' operations included in production rules that have the form of 'IF-THEN' statements. For example, if there is a band in a given wavenumber interval characteristic for a substructure and the peak intensity is within a defined intensity interval, the production rule gives "TRUE" as a result. Typically two to five production rules give this value for a chemical substructure.

*2.4 KNN Classification*

The hitlist obtained by a spectral library search is processed by this method to determine the number of occurrences in the library of each substructure to which the spectrum is classified. The conditional probability for the presence or absence of a substructure are calculated. It is based on the theory for binary classification used previously for mass-spectra [22] and applied by us to IR spectra with LDA and ANNs [15]; the only difference here is that the ANN (or LDA) output value is replaced by the number of hits that contain the corresponding substructure. A description of the algorithm will be given in a separate paper.

*2.5 MCS Analysis*

The maximum common substructure of chemical structures is the largest possible substructure that is present in the given structures. The used software ToSiM [12] contains a tool for the determination of the MCS of two structures represented by two-dimensional connectivity tables. The type of the MCS can be defined by some parameters: usually two sub-structures are con-sidered to be identical (isomorphic) if all atoms (elements) and all bonds (single, double, triple, aromatic) can be matched. Optionally, a further restriction can be applied concerning the number of hydrogen atoms: non-hydrogen atoms are considered to be identical only if the number of hydrogens bonded to them is equal.

The result of applying the MCS concept to the hitlist structures is a list of substructures that are often characteristic for the query compound. This list can be fed into an isomer generator program as a GOODLIST [13, 16].

*2.6 Programming, Hardware Requirements, and Availability*

The program codes of *IRSS* and *IRIS* were written in Delphi, version 2.0. The programs are running under Windows. University scientists and users of non-profit organizations can get a free copy of the executive files and the three libraries via Internet at the address: [http://www.kosnos.com/spectroscopy/iris].

**3 Example of application**

In the example the IR spectrum of 4-phenyl-butanoic acid from ILRIB03 library is considered as unknown. The compound molecular formula is $C_{10}H_{12}O_2$, and using it the software MOLGEN generates 42 252 519 isomers. The spectrum is classified with the aid of LDA and ANN classifiers for 20 chemical substructures, and the library search result (a hitlist) is processed by KNN classifiers for the same 20 chemical substructures and by using the concept of MCS. The results obtained by the different methods are schematically included in Fig. 1.

All seven implemented search algorithms failed to identify the compound when the *CC13484* library was searched: the visual comparison of the spectral curve of the query spectrum with the first 20 hitlist spectra shows significant differences between them. The analysis of the correlation-coefficient hitlist's structures with the aid of the MCS concept gives 18 substructures suggested to be characteristic for the query structure. The first ten are presented in Fig. 2; they consist only of heavy atoms because hydrogens have not been considered in the MCS determination. Eight of them are present in the query structure, and the carboxylic group and benzene ring are correctly predicted. Two substructures are wrong because they contain too long carbon chains. If the fifth and the tenth MCS (the other correct ones are contained in them) are entered into MOLGEN together with the compound molecular formula the program generates only one - the correct - structure. Of course, entering one of the false MCS will discard the correct structure from the list of generated structures but as could be easily seen both contradict the presence of a benzene ring considering the given molecular formula. As the LDA and ANN classifiers predict benzene ring too (see Table 1b) the user can discard the erroneous MCS from the list of found characteristic substructures.

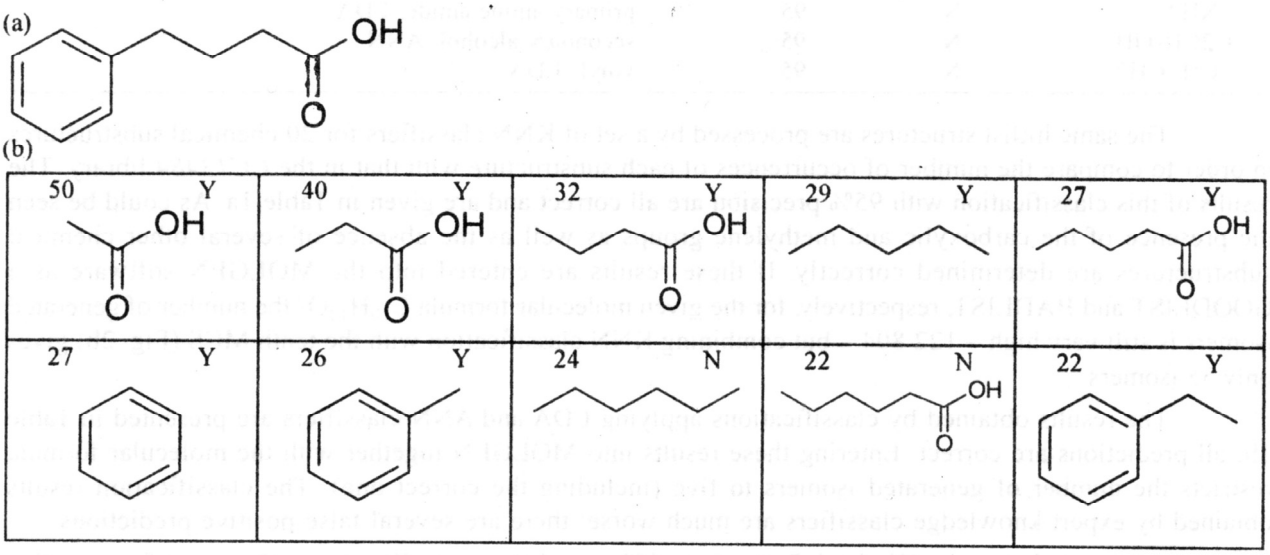


Fig. 2 a) The query structure; b) first ten characteristic substructures found by the MCS analysis of the hitlist structures obtained by full-curve spectral search with correlation-coefficient similarity measure in the CC13484 library; size of the hitlist was 50. Y, substructure is part of the query structure; N, substructure is not part of the query structure; the number given is the frequency (number of hitlist structures containing the substructure).

Table 1. Results from the classification with a threshold precision of 95%. (a) KNN classifiers based on 50 hits from a hitlist obtained by least square full-curve search in the *CC13484* library; (b) LDA and ANN classifiers. *Prec*, estimated precision of classification in percent; substructures were classified to be present in the unknown (marked by "Y") and they were classified to be absent (marked by "N"); all these predictions are correct

(a)

*Interpreted spectrum of 4-Phenylbutyric acid as 'unknown'*

| *Classifier* | *Y/N* | *Prec* | *Comments* |
|---|---|---|---|
| Acid | Y | 99 | Carboxylic acid; KNN LS 50 |
| CH2 | Y | 98 | Methylene; KNN LS 50 |
| CHOH | N | 99 | Secondary alcohol; KNN LS 50 |
| Phenol | N | 98 | Phenol; KNN LS 50 |
| OCH3 | N | 97 | Methoxy; KNN LS 50 |
| *m*-benzene | N | 96 | Meta substituted benzene; KNN LS 50 |

(b)

*Interpreted spectrum of 4-Phenylbutyric acid as 'unknown'*

| *Classifier* | *Y/N* | *Prec* | *Comments* |
|---|---|---|---|
| phenyl | Y | 99 | mono substituted benzene, ANN |
| phenyl | Y | 99 | mono substituted benzene, LDA |
| >CH2 | Y | 97 | methylene group, ANN |
| –COOH | Y | 96 | carboxylic acid, LDA |
| –CH2-OH | N | 99 | primary alcohol, ANN |
| –CH2-OH | N | 99 | primary alcohol, LDA |
| C3C-OH | N | 99 | tertiary alcohol, ANN |
| C3C-OH | N | 98 | tertiary alcohol, LDA |
| *p*-benzene | N | 97 | para substituted benzene, LDA |
| –C(CH3)3 | N | 97 | tertiary butyl, LDA |
| *p*-benzene | N | 96 | para substituted benzene, ANN |
| phenol | N | 96 | phenol, arbitrary substituted; ANN |
| –CH=CH2 | N | 96 | vinyl, ANN |
| *m*-benzene | N | 95 | meta substituted benzene, ANN |

| | | | |
|---|---|---|---|
| –NH2 | N | 95 | primary amine/amide, LDA |
| C2CH-OH | N | 95 | secondary alcohol, ANN |
| –CH=CH2 | N | 95 | vinyl, LDA |

The same hitlist structures are processed by a set of KNN classifiers for 20 chemical substructures in order to compare the number of occurrences of each substructure with that in the *CC13484* library. The results of this classification with 95% precision are all correct and are given in Table 1a. As could be seen the presence of the carboxylic and methylene groups as well as the absence of several other chemical substructures are determined correctly. If these results are entered into the MOLGEN software as a GOODLIST and BADLIST, respectively, for the given molecular formula $C_{10}H_{12}O_2$ the number of generated isomers is still very high – 122 804 – but combining KNN classification with the tenth MCS (Fig. 2b) gives only 35 isomers.

The results obtained by classifications applying LDA and ANN classifiers are presented in Table 1b; all predictions are correct. Entering these results into MOLGEN together with the molecular formula restricts the number of generated isomers to five (including the correct one). The classification results obtained by expert-knowledge classifiers are much worse: there are several false positive predictions.

**Acknowledgements:** We thank R Neudert of Chemical Concepts (Weinheim, Germany) for providing the SpecInfo IR database. We are grateful to the Late J T Clerc and to E Pretsch (ETH Zurich, Switzerland) for making this database available in an appropriate format. We also thank A. Kerber and R. Laue (University of Bayreuth, Germany) for providing the software MOLGEN.

**References**

1. Gray N, *Computer-Assisted Structure Elucidation,* (John Wiley, New York), 1986.
2. Munk ME, *J Chem Inf Comput Sci*, 38 (1998) 997.
3. Luinge H, *Vib Spectrosc*, 1 (1990) 3.
4. Hemmer M, Steinhauer V, Gasteiger J, *Vib Spectrosc*, 19 (1999) 151.
5. Ehrentreich F, *Fresenius J Anal Chem*, 357 (1997) 527.
6. Debska B, Guzowska-Swider B, Cabrol-Bass D, *J Chem Inf Comput Sci*, 40 (2000) 330.
7. Novic M, Zupan J, *J Chem Inf Comput Sci*, 35 (1995) 454.
8. Judge K, Brown C, Hamel L, *Anal Chem*, 80 (2008) 4186.
9. Clerc J, in *Computer-Enhanced Analytical Spectroscopy*, Meuzelaar H, Isenhour T (eds) ,(Plenum Press, New York), 1987, pp145 - 162.
10. Penchev P, Sohou A, Andreev G, *Spectrosc Lett*, 29 (1996) 1513.
11. Penchev P, Kochev N, Andreev G, *Compt Rend Acad Bulgare des Sci*, 51 (1998) 67
12. Scsibrany H, Varmuza K, in, *Software Development in Chemistry*, Jochum C (ed), (Gesellschaft Deutscher Chemiker, Frankfurt am Main), 1994, vol 8, pp 235-249.
13. Varmuza K, Penchev PN, Scsibrany H, *J Chem Inf Comput Sci*, 38 (1998) 420.
14. Chemical Concepts, PO Box 10 02 02, D-69442 Weinheim, Fed Rep of Germany.
15. Penchev PN, Andreev GN, Varmuza K, *Anal Chim Acta*, 388 (1999) 145.
16. Varmuza K, Penchev P, Scsibrany H, *Vib Spectrosc*, 19 (1999) 407.
17. Buchanan B, Sutherland G, Feigenbaum E, in *Machine Intelligence*, Meltzer B, Michie D (eds), (Edinburgh University Press, Edinburgh), vol 4, 1969, pp 209-254.
18. Benecke C, Grund R, Hohberger R, Kerber A, Laue R, Wieland T, *Anal Chim Acta*, 314 (1995) 141.
19. Penchev P, *Application of Chemometric Methods for Identification of Organic Compounds from Their Infrared Spectra*, Ph D Thesis, Plovdiv, Bulgaria, 1998.
20. Pretsch E, Clerc T, Seibl J, Simon W, *Tables of Spectral Data for Structure Determination of Organic Compounds*, (Springer, Berlin), 1989.
21. Andreev G, Argirov O, Penchev P, *Anal Chim Acta*, 284 (1993) 131.
22. Varmuza K, Werther W, *J Chem Inf Comput Sci*, 36 (1996) 323.

[*Received*: 27.03.2009; *accepted*: 16.04.2009]