Software Approaches for Structure Information Acquisition and Training of Chemistry Students

Nikolay T. Kochev, Plamen N. Penchev, Atanas T. Terziyski,
George N. Andreev

Department of Analytical Chemistry, University of Plovdiv,
Tsar Assen Str. 24, 4000 Plovdiv, Bulgaria

Abstract
The process of determination of chemical structures is considered in respect to its software implementation. Several main stages are distinguished and modeled. A proper software implementation is proposed for each stage. The stages represent important scientific tasks as well as basic steps in the training of chemistry students. A spectral database management software is developed to be the main software module. Additional modules are included to realize the other stages using pattern recognition methods, expert knowledge based methods and statistical methods.

Key words: *spectral database, spectra interpretation, structure determination, binary classifier, expert knowledge*

Introduction

The applications of contemporary software systems in the field of scientific research and students education have always been of interest. The study of chemical compounds and most of the other applications related to chemistry research demand the most advanced computer and software technologies. The first tree numerical (non bibliographic) databases introduced on-line in 1971 were chemical in nature [1]. Up to date there are more than 15 millions registered chemical compounds. Usually, the largest industrial or scientific databases contain information for less than 200 000 compounds, therefore the approach for direct identification of an unknown compound by means of search in a chemical database in most of the cases does not give the desired result; the probability even the biggest database to contain the unknown compound is too small. On the other hand the problem of structure determination is very important in many research fields and also in the process of education in natural sciences. The most effective and cheapest way for acquisition of information for the structure of an unknown compound is the implementation of one or several spectroscopic techniques and consequent interpretation of the registered spectra [2, 3]. For the scientist and student, the process of spectra interpretation requires a lot of skills, expert knowledge and accuracy in order to extract the structural information from the spectral data. This article depicts several software approaches for

acquisition of structure information on the base of the spectral data. Each software method models and implements a concrete task performed by the scientist or student. The developed by us software model is applied in our department for education purposes and in our and other laboratories for studying of chemical compounds.

Acquisition of Structural Information

One of the main problems of analytical chemistry is to determine the structure of an unknown compound on the base of data obtained by analytical experiment or spectroscopic techniques. As it is well know the spectrum can be regarded as a function of the structure.

$$Spectrum = f(Structure) \quad (1)$$

the *Structure* is presented by a two dimensional chemical graph; *Spectrum* typically is represented by data arrays. Clerc showed the principal impossibility to obtain the reverse function, f^{-1} [3]. It is possible, however, to obtain partial information for the structure by means of various methods, which combined in a proper way could generate the query structure or at least fragments of it. Figure 1 summarizes the process of structure determination:

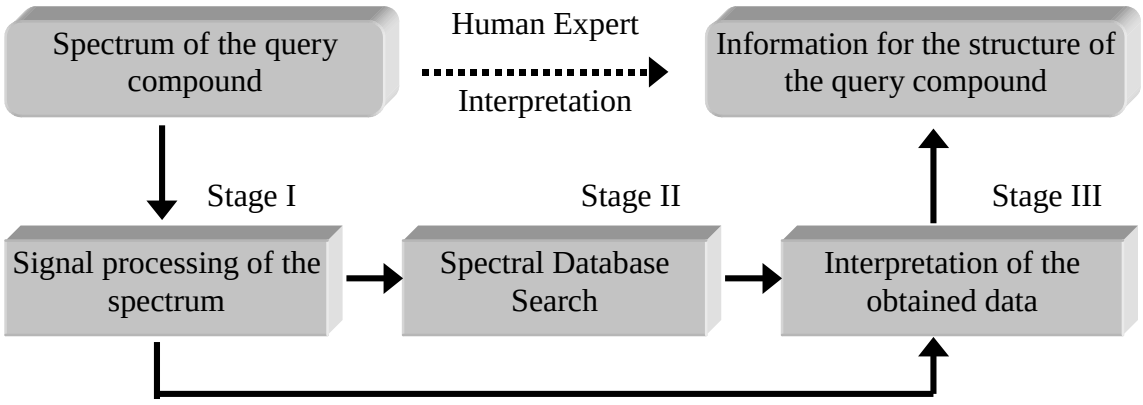

Figure 1. The main stages of the developed system for structure determination.

The interpretation of spectral data can be performed in two ways: with or without use of computer. In both cases at least three stages are distinguished: processing of the rough signals, search in spectral databases, and interpretation or evaluation of results obtained in the previous stages. Even if the computer is not used (denoted by a dashed arrow in fig. 1), these stages are also performed but mentally and solely depending on the skills of the scientist. Some of the main factors that imply the efficiency of software realization of the described stages are: *(i)* each stage can be mathematically

modeled and presented as well defined procedure; *(ii)* some stages require a lot of data manipulations or calculations that are quite ineffective or impossible to be done by hand; *(iii)* the logic of the human expert can be appropriately presented in a knowledge base or simulated by pattern recognition methods. On the other hand these stages are important when students are being trained to interpret spectral data. The developed software modules allow educator to focus students efforts on the logic and the practical use of the expert knowledge, rather than the technical details and concrete calculations which are not the primary goal in the training of chemistry students.

Since there is principal impossibility to obtain the invert function of f (eq. 1), the interpretation of spectral data is based on a number of correlations [2] between structural fragments and different spectral features. These correlations could be regarded as models of some restrictions of the invert function, f^{-1}. There is a variety of spectral features, some are directly derived from the spectral data (fig. 1, the case when Stage II is omitted), while other features are derived from the spectral database search results. Some of the used spectral features are described bellow.

Signal Processing of Spectra (Stage I)

Generally the spectra describe how given physical parameters functionally depend on the change of other parameters while a specific physical process is performed. The presumption is that this dependencies are specific for each chemical compound, hence on the basis of spectral data structures are determined. There is a variety of spectroscopic techniques, MASS, infrared, ultraviolet-visible, Raman etc. In each case the spectra are manipulated as signals registered by appropriate technical equipment. Typically the spectra are one dimensional signals presented in digital form as data arrays; there are some two dimensional applications as well. Since the spectral data are crucial in respect to the structure determination, the spectra signals should be processed to allow maximal gain of structure information. Spectra have specific components, spectral bands for example, that are the main source of structural information. The aim of the signal processing is to make these spectral components distinguishable. The basic routines applied to the spectrum before its actual use are noise reduction, smoothing and base line correction, usually done automatically. Deconvolution [8] of the signals is another important procedure used in the software approach. Figure 2 demonstrates the use of software, PSD, to process an example spectra. PSD (Program for Spectra Deconvolution) is a Windows™ based software [8], developed by us to perform signal processing of infrared, Raman and UV-visible spectra.

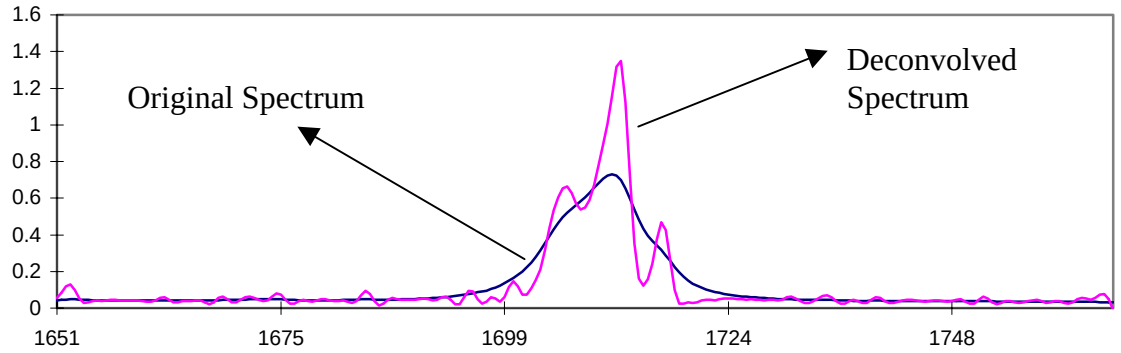

Figure 2. The graphics of original and deconvolved spectra; the processed by PSD spectrum contains narrow bands with higher intensity;

The original spectrum band (see fig. 2) is decomposed to three intense bands, which implies additional information for the structure (each band corresponds to a particular structural fragment). The expert would distinguish the three overlapped bands without software deconvolution by observing the inflection points near the band maximum of the original spectrum. However the student most probably would not decompose correctly this peak to three bands and thus would not gain the best structure information.

Search in Spectral Databases (Stage II)

Spectral database contains records for a number of chemical compounds. Usually each record consists of several components: spectrum, peak table, structure, IUPAC name, brutto formula and additional data for some chemical and physical properties. Database is divided into logical units called spectral libraries, where each library contains information for a given set of chemical compounds with registered spectra for a particular spectroscopic method. The basic feature of such database is the library spectral search using mainly the spectral data. All other kinds of searches typical for every relational database management system are not directly used in the process of structure determination. However, they are of great help in the process of knowledge base building and for statistical characterization of the pattern recognition methods described in the next section. As it was mentioned above, the probability that the spectral database would contain a spectrum corresponding to the studied unknown compound is too low, therefore one can not expect direct identification of the studied compound. Nevertheless, these database searches generate enough information which manipulated in a proper way could help researcher determine the unknown structure or at least fragments of it. The spectral libraries searches are based on the presumption that a similar chemical compounds exhibit similar spectra [3]. In order to perform library search, an appropriate similarity measure is

needed. Basically there are two kinds of spectral similarity measures: full spectra curve measures and peak table based measures. The software IRSS (InfraRed Search System) has been developed in our laboratory and discussed in previous publication [7]. It is a Windows based program coded in Object Pascal; the source contains more than 35 000 lines.

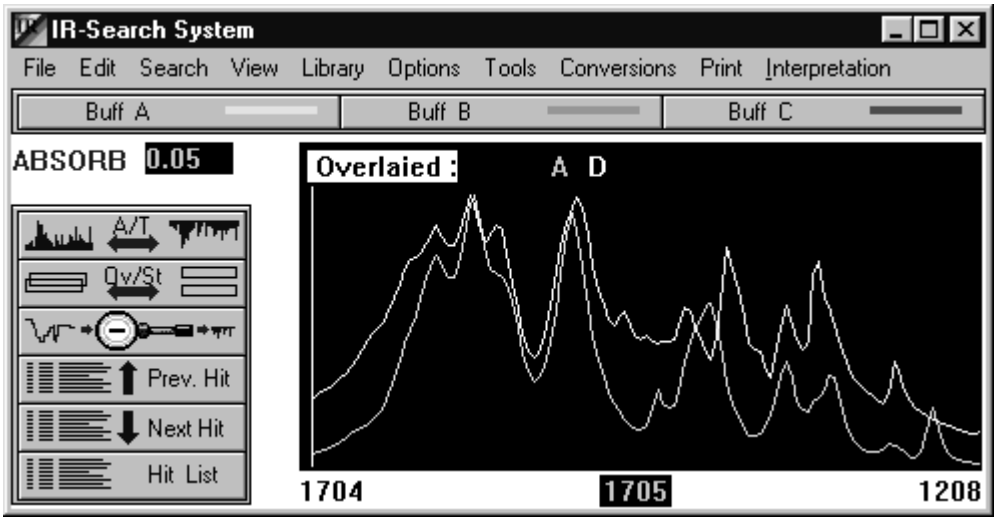

Figure 3. The resized main window of the software IRSS. The spectrum of the investigated compound and a search result spectrum are overlaid.

IRSS implements four types of full spectral curve similarity measures using Euclidean distances, scalar product and correlation coefficient of two spectra. Additionally IRSS supports three types of peak table similarity measures based on counting the number of matched peaks for a pair of spectra. The main result of the spectral search is the so called “hitlist” which contains the compounds most similar to the investigated one. The proper evaluation of the hitlist structures gives information relevant to the studied compound; this is described in the next section.

Substructure search is a new recently developed feature of IRSS. The procedure is based on the so called structure isomorphism check i.e. whether a query structure fragment is a substructure of another structure. This way database is searched for a specific structural fragments. The latter search allows students manually to check or prove correlations between structural fragments and corresponding spectra as well as variety of routines used during the spectra interpretation.

Software Approaches for Spectra interpretation (Stage III)

The actual acquisition of structure information is carried out in Stage III (see fig. 1) by means of several methods implemented by us as separate software applications or additional modules of IRSS. Most of the methods

apply the concept of binary classifier. Binary classifiers (see fig. 4) are procedures performed for different structural fragments. Each classifier has as an input, the features (parameters) derived from the spectroscopic data [5] and its output data determines whether the unknown chemical structure contains the structural fragment or not. Additionally the results reliability is characterized by a probability and explanations.

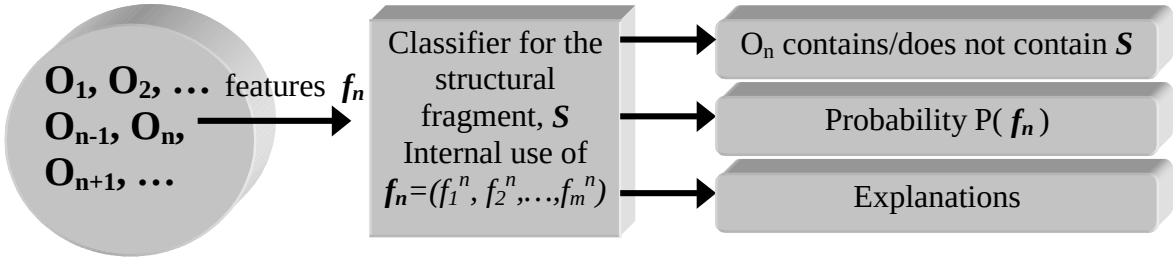

Figure 4. The implemented by us concept of binary classifier; the classifier determines the presence or absence of a particular structural fragment; **U**={O_1,O_2,…,O_n,…} is the space of all chemical objects; vector, $\boldsymbol{f_n}$, represents the features of the compound O_n

Each classifier divides the space of all chemical objects into two classes: *class 0* consists of compounds which do not contain the substructure and *class 1* are those compounds that contain the substructure. The features calculation and the probabilistic characterization depend on the classification methods. Some of the main approaches for binary classification are described as follows:

1. *k nearest neighbors method (kNN).* It is mathematically simple non-parametric multi-class pattern recognition method, based on so-called 'majority vote' procedure [4]. Using spectral library search, the k nearest neighbors of the unknown compound, x, are determined (see fig 5). Classification for a given substructure, S_i, is done on the base of the number of *class 1* objects, n_i, among the structures from the hitlist. For each classifier, the probability function, $P(n_i)$, is statistically characterized by a training set of structures (the used feature in this method is n_i). We have realized software module for automatic generation of kNN classifiers. The generator uses specially designed by us Structure editor, applied for structure search as well.

2. *Linear Discriminant Analysis (LDA)*. Using a learning set of objects, a hyper-plane is obtained to divide the space **U** into two subspaces corresponding to *class 0* and *class 1* respectively [4]. The unknown compound, x, is classified in accordance with the subspace it is placed in. LDA method together with the artificial neural networks classifiers are implemented in an additional software module IRIS.

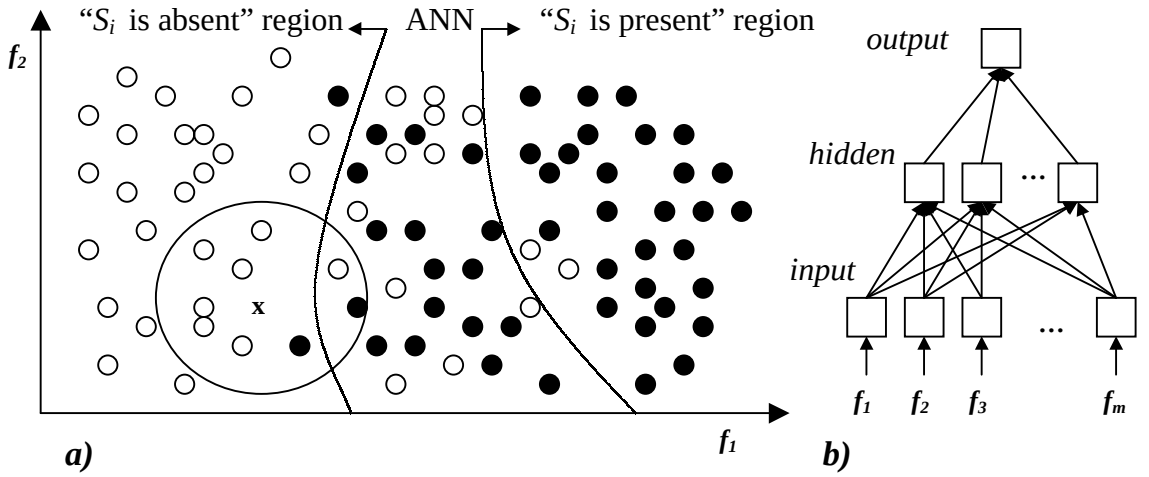

Figure 5. a) kNN and ANN methods realization; the space of chemical objects, **U**, is represented by the features f_1 and f_2. kNN method classifies, x, to be in *class 0*, since amidst the 9 nearest neighbors (k=9), 7 does not contain the substructure S_i. ANN also classifies x to be in *class 0* because x is in the *class 0* region. b) The realized by us ANN consists of 3 layers of neurons: input, hidden and an output neuron.

3. *Artificial Neural Networks (ANN)*. Non-liner modeling of the classes is obtained. Analogously to the LDA, two regions of **U** are obtained (fig. 5). The region surfaces are defined by the coefficients of the ANN [5]. Each ANN has a single output neuron since a binary classification is performed.

4. *Maximum Common Substructure Approach (MCS)*. The structures from the hitlist are topologically characterized, searching the maximal commons substructure of each pair of structures and taking the most frequently occurring MCSs [6]. MCS method is integrated in IRSS.

5. *Expert knowledge based system*. A knowledge base is built implementing heuristics derived from spectra-structure correlations. The inference engine is realized in the software EXPIRS [9]. This is a program coded in PROLOG, which interprets the peak table of the infrared spectrum of an unknown compound.

Example of Application

To illustrate the above approaches, compound "Anthranilic acid, methyl ester" was regarded as unknown and its structure was determined on the basis of its infrared spectrum. The compound was searched in a spectral database with 1000 compounds. After that 40 ANN/LDA and 20 kNN binary classifiers were applied as well as MCS analysis and interpretation with the help of the expert system. Figure 7 summarizes the results of the combined software application of IRSS, IRIS and EXPIRS. Each method has advantages and some disadvantages, therefore, the combination of all methods exhibited the best results.

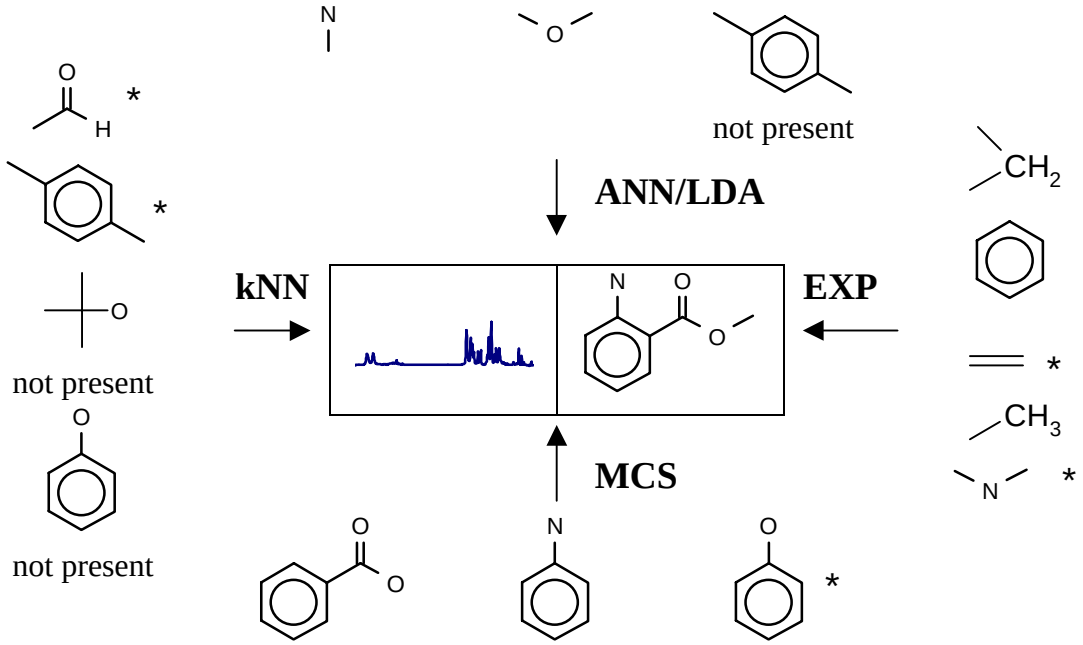

Figure 6. Combined approach for structure determination. The “unknown” compound and its IR spectrum are in the center; the obtained structural fragments are placed on the left, right, top and bottom respectively. The erroneous results are denoted with ‘*’.

References

[1] Meyer, E.; Funkhouser, N. F. A Brief History of Networkign in the U.S. *J. Chem. Inf. Comput. Sci.* **1998,** *38, 951-955*.

[2] Munk, M. E. Computer based structure determination: Then and now. *J. Chem. Inf. Comput. Sci.* **1998,** *38, 997-1009*.

[3] Clerc, J.T. Automated Spectra Interpretation and Library Search Systems, pp. 145-162 in : H.L.C. Meuzelaar, T.L. Isenhour (eds.); *Computer-Enhanced Analytical Spectroscopy.* Plenum Pres, New York, **1987**.

[4] Massart, D.L.; Vandeginste B.G.M.; Deming S.N.; Michotte Y.; Kaufman L. *Chemometrics: A Textbook.* Elsevier, Amsterdam, **1988.**

[5] Penchev, P.; Andreev, G.; Varmuza, K. Automatic Classification of Infrared Spectra Using a Set of Improved expert-based features. *Analytica Chimica Acta*. **1999**, 388, *145-159.*

[6] Varmuza, K.; Penchev, P.N.; Scsibrany, H. Maximum common substructures of organic compounds exhibiting similar infrared spectra. *J. Chem. Inf. Comput. Sci.* **1998**, 38, *420 - 427*.

[7] Penchev, P.N.; Kochev, N.T.; Andreev, G.N. IRSS: A program system for infrared library search. *Comp. Rend. Acad. Bulg. Sci.* **1998,** 51, 67-70.

[8] Kochev, N.T; Rogojerov, M.I; Andreev, G.N. A new graphical approach for improved user control of Fourier self-deconvolution of infrared spectra, *Vibrational spectroscopy*. **2001**, 25 (2),*177-183.*

[9] Andreev, G.N.; Argirov, O. K. EXPIRS, an expert system for generation of alternative sets of substructures, derived by of infrared spectra interpretation. *Analytica Chimica Acta*, **1996**, 321, *105-111*.