Automatic Generation of Knowledge Base from Infrared Spectral Database for Substructure Recognition

Barbara Dębska,*,† Barbara Guzowska-Świder,† and Daniel Cabrol-Bass‡

Department of Computer Chemistry, Rzeszów University of Technology, 35-041 Rzeszów, Poland and GRECFO—LARTIC, University of Nice Sophia-Antipolis, Nice, France

Received August 3, 1999

This paper presents a new methodology of chemical substructure recognition by interpretation of an infrared spectrum. The approach in spectrum interpretation is based on the determination of functional groups, which may be present or absent in compounds whose structure is unknown. The process of searching for spectrum−substructure correlation is realized by application of a statistical algorithm. In this method, correlations are generalized and condensed into a set of interpretation rules which are applied to the interpretation of an unknown compound's spectrum in order to predict whether the respective substructures are present or absent in the unknown molecule.

1. INTRODUCTION

Over the past 25 years, numerous attempts have been made[1,2] to improve the speed and the reliability of the structure elucidation process with the aid of computers.

Library Search Methods. The first group of programs to be developed for computerized spectra interpretation was based on the retrieval of the spectra of the unknown compound in spectral databases. Because IR spectra are subject to noise and to experimental conditions, exact matches cannot be expected. Therefore, the algorithms usually compute a degree of similarity between the spectra of the unknown and the spectra in the database to construct a list of compounds sorted by decreasing order of spectral similarity.[3−7] These programs need a large database of high quality and cannot solve the problem when the spectrum of the unknown compound is not stored in the database. However, it seems reasonable to think that similar spectra retrieved correspond to compounds of structure more or less similar to the unknown.

Correlation Tables and Expert Systems. A different type of approach in spectra interpretation is based on the use of the functional dependencies of the frequency upon definite structural fragments in the form of spectrum−structure correlation tables to determine the substructures that may be present in an unknown compound: CASE,[8] PAIRS,[9−11] SEAC,[12] STREC,[13,14] CHEMICS,[15,16] ESSESA,[17] EXPIRS,[18,19] X-PERT,[20] SPEKTREN,[21] EXPEC,[22,23] ARGIS,[24] etc.[25−29] Each of this type of system solves the problem of substructure recognition by its own algorithm. The first systems[8−12] followed closely the classical ways chemists identify unknown compounds. The table-driven procedure was applied for fast interpretation of infrared spectra and to receive a list of substructures that were probably a part of an analyzed compound.[16,17] In most systems, the knowledge about spectrum−structure correlation is formalized as interpretation rules. Systems in which interpretation rules are not an integral part of the interpretation program but are gathered in a separate file (rule base or knowledge base) which is processed by a specific code (called an inference engine) to draw conclusions are described as having an "expert system" architecture. All the above-mentioned programs for structure elucidation used the spectrum−structure correlation taken from the literature, directly stated by human experts, or extracted automatically from spectral databases. Affolter et al.[30] has shown that direct use of tabulated normal frequency ranges taken from the literature[31] lead to a high percentage of cases (50% as mean) in which at least one of the ranges was empty of peaks. This fact calls for the development of more sophisticated methods for interpretation of infrared spectra.

Pattern Recognition and Neural Networks. Another approach which has been explored for computer-aided structure elucidation includes various cluster analysis and pattern recognition methods.[32,33] These methods make use of spectral databases in the learning and testing phases to develop a computerized model for inferring the presence or absence of the respective structural fragments in the unknown substance from its spectrum. Progress in the theory and practice of artificial neural network (ANN) enables the application of ANN in analytical chemistry for solving spectrum−substructure correlation problems.[34−44] ANN programs are used both for substructure determination from spectra and for spectra simulation from structure.

Knowledge Acquisition. In recent years, automatic knowledge acquisition from databases has become the subject of intensive research in data processing. As a matter of fact, a vast amount of information and knowledge is embedded in large chemical databases nowadays available. Therefore, methods for revealing information and constructing knowledge automatically from factual databases are of considerable interest.

The method originates from database theory, statistical methods, machine learning, rough sets theory, and other domains related to databases and artificial intelligence.[45,46]

[*] Corresponding author: E-mail: bjdebska@prz.rzeszow.pl.
[†] Rzeszów University of Technology.
[‡] University of Nice Sophia-Antipolis.

10.1021/ci990271x CCC: $19.00 © 2000 American Chemical Society
Published on Web 03/03/2000

Table 1. Part of the List of Substructures Identified by the System SCANKEE

no.	substructure	substructure environment	comments
1	CH_3-		
2	CH_3-	C	$C \Rightarrow -CH_2-, -CH<, >C<$
3	CH_3-	$-O-$	
⋮			
19	$-CH_2-$		
20	$-CH_2-$	C, $-O-$	$C \Rightarrow -CH_3, -CH_2-, -CH<, >C<$
21	$-CH_2-$	C, arom ring	$C \Rightarrow -CH_3, -CH_2-, -CH<, >C<$
⋮			
51	$-OH$		
52	$-OH$	C	$C \Rightarrow -CH_3, -CH_2-, -CH<, >C<$
⋮			
67	$-NH_2$		
⋮			
117	$-CO-$	arom ring	
⋮			
131	$-COOH$		
132	$-COOH$	C	$C \Rightarrow -CH_3, -CH_2-, -CH<, >C<$
133	$-COOH$	arom ring	
⋮			
183	$CH_2{=}CH-$		
⋮			
195	pyridine ring		
196	benzene ring	isolated, uncondensed	
197	monosubstituted benzene ring		
198	1,2-disubstituted benzene ring		
⋮			
215	naphthalene ring	isolated, uncondensed	
⋮			

However, much knowledge can be obtained from data without the use of sophisticated techniques or tools. A large number of essential facts are hidden in the form of difficult-to-determine links. In such cases, the traditional statistical methods, e.g., regression analysis, fail creating a need for tools and techniques which would enable us to find and analyze the structural relationships in the base. They can also be used to identify the relationships, essential causes of facts, and to present the knowledge in the form of general mathematical expressions, sets of prediction rules, decision trees, mathematical models, etc.

In some specific domains, such as economy or sociology, knowledge acquisition techniques have been successfully applied not only to confirm known facts but also to reveal new unknown facts. In molecular spectroscopy, the relationships within a molecular spectral database are often known. The application of the knowledge discovery algorithms was aimed to confirm and eventually to refine facts described in the literature by spectrum−structure correlation tables.

Aims and Goal. Methods of substructure identification discussed above do not lead directly to the complete structure of the unknown. They provide constraints which are used during the next steps of the structure elucidation process which consist in the generation of candidate structures. The reliability of the constraints inferred from spectra in the first steps is of paramount importance for the global performance of the whole system. Any error in the fragments considered as being definitively present or absent in the unknown structure will lead to the generation of erroneous structure, and even worse will prevent the construction of the correct solution. Therefore, the above-mentioned methods of inferring structural information from spectral data are continuously developed. These methods concern spectra processing[47−49] as well as spectra interpretation.[18−20,24,25,50−55] In contrast to earlier approaches based on preestablished correlation tables which lead to binary answers, recent methods focus on the degree of confidence of the answers they produce.

In this paper, we present an improved method for constructing spectrum−substructure correlation, by automatic knowledge acquisition applied to an IR spectra database. The result of this process is a set of rules which are used by an expert system for inferring the possible presence or absence of substructures in an unknown chemical compound from its infrared spectra. Although much work has been done on automatic extraction of interpretation rules from spectral data, we intended on developing new solutions to improve their effectiveness. Our results are compared with those obtained independently[41] on the same collection of spectra using hierarchical feed forward multilayer ANN.

2. METHOD

Structure of the Example Base. Before the algorithm for rule generation from the database could be presented, the *structure of records* composing the example base containing the molecular spectra, and the basic concepts, should be defined. They are as follows:

(a) The F set of object fragment: $F = \{f_1, f_2, ..., f_n\}$

Each element of the set represents a fragment of an object's structure determined unequivocally. In our complete system F is the set of 222 fragments (substructures) which have been investigated (Table 1).

(b) An O object: $O = \{ f_i, f_j, ..., f_k\}$ (for each molecular structure) is a set of fragments selected from the set F.

In the source structure−spectral database, the structure (i.e., the type of atoms, the way of their connection, the type of chemical bonds) of each chemical compound is coded as a connectivity matrix. In the discussed algorithm, an O object represents only a set of these structural fragments which are simultaneously a part of a molecule and belong to the F set

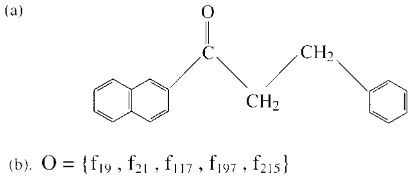


Figure 1. Representation of a chemical compound (a) as an *O* object (b).

of fragment active in infrared spectroscopy. For example, the compound in Figure 1 is represented by object $O = \{f_{19}, f_{21}, f_{117}, f_{197}, f_{215}\}$. The fragment numbers are given in the first column in Table 1. The structure of each f_j fragment is stored as a connectivity matrix, and this code form is used to check if the fragment is a part of the compound structure as stored in the structure−spectral database.

(c) The *OC* set of object features: $OC = \{(co_1, io_1), (co_2, io_2), ..., (co_m, io_m)$, where (co_i, io_i) are the values *describing* an "*i*" feature ($i = 1, 2, ..., m$).

In our case the object features are spectral parameters of absorption bands, i.e., band location (cm $^{-1}$) "co_i" and band intensity (% Transmittance) "io_i".

(d) The *P* example: $P = (O,OC)$.

The *O* object and the *OC* set of its features form one *P* example stored as one record in the database. An example of such a record (for a chemical compound: *β*-cyclooctatetraenylpropionic acid) defined in this way in the database can be a pair (*O*,*OC*) described in the following terms:

$$(O = \{f_{19}, f_{131}\}, \quad OC = \{(2993,14.4), (2599,52.4), (1700,6.1), ..., (669,61.4)\})$$

(e) Example base *BP*: $BP = \{P_1, P_2, ...\}$ is a set of *P* records.

The above-defined example base is a source of hidden knowledge, whose acquisition can be realized through various methods, e.g., statistic method,[21,24,56] pattern recognition method,[32,33] or neural networks.[34−42]

Algorithm of Knowledge Acquisition from an Example Base. In the present work the basic version of AQ algorithm (*A*lgorithm *Q*uasi-optimal)[57] was improved. This algorithm makes use of the following data:

POS set: set of positive examples

NEG set: set of negative examples

LEF correlation vector (named by Michalski[57] *L*exicographical *E*valuation *F*unctional).

The knowledge acquisition algorithm is applied for each selected fragment f_j belonging to the *F* set. The POS_j set, NEG_j set and LEF_j correlation vector are created separately for each f_j fragment. The POS_j (NEG_j) set includes those *P* examples from the *BP* example base, which contains (does not contain) f_j fragment for which a LEF_j correlation vector is searched.

The LEF_j correlation vector is defined as

$$\mathrm{LEF}_j = \{[(c_1,tc_1),(i_1,ti_1)], ..., [(c_k,tc_k),(i_k,ti_k)]\}$$

where in (c_m,tc_m) c_m is a value of an *m* feature (frequency) and tc_m is a tolerance interval for c_m feature, in (i_m,ti_m) i_m is a value of an *m* feature (intensity) and ti_m is a tolerance interval for i_m feature, and for $m = 1, 2, ..., k$, *k* is a number of spectral features found as characteristic for f_j fragment by algorithm AQ.

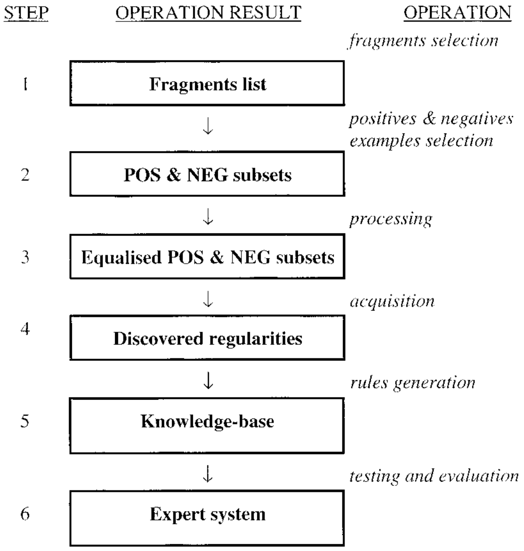


Figure 2. Algorithm of knowledge acquisition from spectral database.

The scheme of the developed algorithm for knowledge acquisition from the BP example base, is shown in Figure 2. The BP base was transformed into a knowledge base in the following steps:

1. Selection of f_j fragments ($j = 1, 2, ..., n$) for which the knowledge base will be created.

2. Partition of the *BP* base into POS_j and NEG_j subsets.

3. Elimination (by a random process) of examples from POS_j or NEG_j sets to obtain the same number of positive and negative examples.

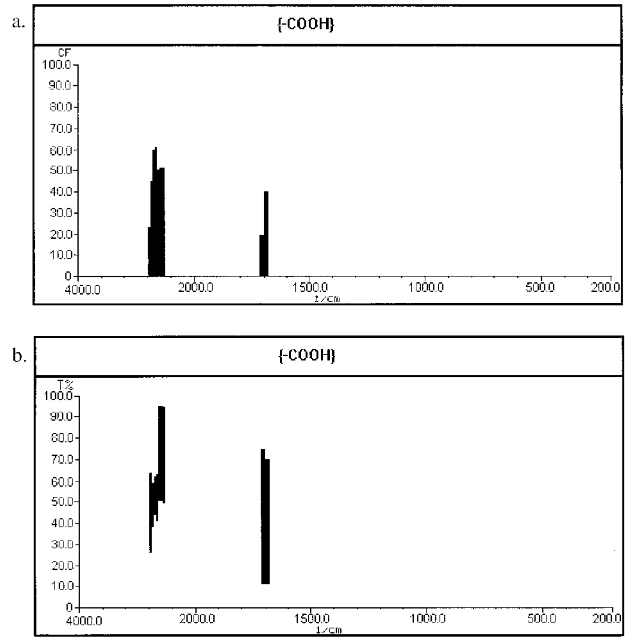
4. Creation of LEF_j correlation vectors by analysis of POS_j and NEG_j sets.

5. Generation of a rule knowledge base consistent with regularities (represented by LEF_j correlation vectors) discovered in the example base.

6. Test of the rule knowledge base using the inference engine (a SEK module of the SCANKEE[58] system) and a set of objects for which structure and features are known.

The construction of the LEF_j correlation vector in step 4 is achieved in two phases. First, the number of spectral features to be retained and their frequency are fixed; second, their associated intensity range is determined. In the first phase, the complete IR spectrum is divided into 140 intervals: 50 intervals of 40 cm^{-1} width in the 4000−2000 cm^{-1} region and 90 intervals of 20 cm^{-1} width in the 2000−200 cm^{-1} region. A $V_{\mathrm{POS},j}$ vector is calculated from a POS_j set. The component values indicate the number of spectra with bands falling in the appropriate interval. In the same way a $V_{\mathrm{NEG},j}$ vector is calculated from a NEG_j set. The substructure diagnostic vector $V_{\mathrm{D},j}$ is defined as the difference between $V_{\mathrm{POS},j}$ and $V_{\mathrm{NEG},j}$. Next, $V_{\mathrm{D},j}$ is "normalized" by dividing each of its components by the cardinality of POS_j and NEG_j sets leading to the CF_j vector. Each component $CF_{j,i}$ ($i = 1, ..., 140$) can be interpreted as a confidence factor of the corresponding spectra interval *i* for the considered fragment *j*. The LEF_j vector is built from the CF_j vector by rejecting the components whose values are below a given threshold value (Figure 3a).

In the second phase, the intensity range (i_m, ti_m), for each component m of the LEF_j vector is determined by seeking



c.

{-COOH}

Correlation table:

No	Band location		Band intensity		CF
	[1/cm]	[1/cm]	[T%]	[T%]	[%]
1	2680	2640	40.36	63.84	60.29
2	2720	2680	42.36	61.19	59.00
3	2600	2560	50.23	94.84	51.02
4	2560	2520	48.64	94.41	51.00
5	2640	2600	50.23	94.84	49.22
6	2760	2720	38.46	60.84	46.82
7	1700	1680	10.00	70.13	39.92
8	2800	2760	26.92	64.49	23.98
9	1720	1700	10.00	75.01	19.69

Figure 3. Spectrum−structure correlation generated for −COOH substructure. (a) CF_i values of LEF_{COOH} vector (CF_i > 18%); (b) intensity intervals of LEF_{COOH} vector; (c) values of LEF_{COOH} components (band location) and intensity intervals (band intensity).

the minimum and maximum values of transmittance in spectra of the POS_j set.

An example of LEF_j correlation vector for the −COOH fragment is graphically depicted in Figure 3b. The components of the LEF_{COOH} ordered according to decreasing values of CF_j factors form the correlation table, which is presented in Figure 3c. For this substructure the LEF_{COOH} vector has nine components, where c_m is the center of a frequency interval and tc_m is 20 or 10 (cm^{-1}). Contiguous components form two separate diagnostic regions. The intervals 1−6 and 8 form the first diagnostic region 2800−2520 cm^{-1} while intervals 7 and 9 make the second one, 1720−1680 cm^{-1}.

The discussed step 4 of the algorithm (determination of LEF_j vector and building the correlation table) is realized for all selected f_j fragments (j = 1, 2, ..., n). Next, the interpretation rules are generated (step 5) and the rule knowledge base is created.

Automatic Generation of a Rule Knowledge Base. The aim of the knowledge acquisition algorithm is to answer the question, "Which spectral features, absorption band parameters in the IR spectrum, result from the presence of the analyzed substructure in the molecule?" Fixing these relationships, initially as LEF_j correlation vectors, afterward as prediction rules, gives the possibility for generation of a rule knowledge base. This knowledge base with an inference engine forms an expert system for substructure recognition of chemical compounds from infrared spectra.

The simplest form of a production rule is IF <conditions> THEN <hypothesis>.

In the earlier version of our system[50] the standard format of f_j substructure identification rule was as follows:

```
RULE
IF    band_location ∈ interval_frequency[j, jr] AND
      band_intensity ∈ interval_band_intensity[j, jr]
THEN substructure fj IS present with partial_probability_index [j, jr]
```

where j_r is the number of the diagnostic region characteristic for the considered substructure and the partial_probability_index[j,j_r] = 1/j_r.

In the new version of the algorithm the rule format is modified and the partial_probability_index calculation takes into consideration the values of $CF_{j,i}$ factors. The actual format of the interpretation rules is as follows:

```
RULE
IF
  ( (band_location ∈ interval_frequency[j,1,1] AND
       band_intensity ∈ interval_band_intensity[j,1,1])
   OR
     (band_location ∈ interval_frequency[j,1,2] AND
       band_intensity ∈ interval_band_intensity[j,1,2])
   OR
   . . .
     (band_location ∈ interval_frequency[j,1,j1max] AND
       band_intensity ∈ interval_band_intensity[j,1,j1max]) )
THEN substructure_j IS present with partial_probability_index [j,j1]
. . .
RULE
IF
  ( (band_location ∈ interval_frequency[j,jr,1] AND
       band_intensity ∈ interval_band_intensity[j,jr,1])
   OR
     (band_location ∈ interval_frequency[j,jr,2] AND
       band_intensity ∈ interval_band_intensity[j,jr,2])
   OR
   . . .
     (band_location ∈ interval_frequency[j,jr,jrmax] AND
       band_intensity ∈ interval_band_intensity[j,jr,jrmax]) )
THEN substructure_j IS present with partial_probability_index [j,jr]
RULE
probability_index [j] = sum of partial_probability_indexes [j,j1] ... [j,jr]
```

where j_r is the number of the diagnostic regions characteristic for the f_j analyzed fragment, $j_{1\max}$ is the number of intervals in the first diagnostic region, and $j_{r\max}$ is the number of intervals in the rth diagnostic region.

The partial_probability_index[j,m] (m = 1, ..., j_r) for the mth diagnostic region depends on its arithmetic mean $CF_{j,m}$

Table 2. Spectrum−Structure Correlation Generated for Analyzed Substructure

no.	sub-structure	spectral region (cm^{-1})	intensity region (% transmittance)	threshold values of CF vector (%)
1	hydroxylic	3400−3360	10.0−74.5	40
		3360−3320	10.0−94.8	
2	carboxylic	2800−2760	55.7−94.8	20
		2760−2720	38.5−94.7	
		2720−2680	42.4−94.1	
		2680−2640	40.4−94.8	
		2640−2600	50.2−94.7	
		2600−2560	41.6−94.7	
		2560−2520	48.6−94.4	
		1720−1700	10.0−75.3	
		1700−1680	10.0−74.4	
3	amino	3400−3360	15.3−94.8	18
		3360−3320	27.0−94.7	
		3320−3280	10.0−94.4	
		3280−3240	10.1−94.8	
		3240−3200	15.3−94.8	
		1640−1620	10.0−91.0	
		1620−1600	10.1−94.8	
		1300−1280	16.0−85.2	
4	benzenic	3120−3080	20.0−94.8	28
		3080−3040	22.3−94.2	
		1620−1600	10.0−94.3	
		1520−1500	10.0−94.0	
		1500−1480	10.0−94.4	
5	ethylenic	3080−3040	34.9−94.2	28
		3040−3000	42.0−93.3	
		1660−1640	18.8−92.9	
		1640−1620	10.0−91.0	
		1000−980	10.0−89.3	

calculated for all intervals in the *m*th region:

$$\text{partial\_probability\_index}[j,m] = CF_{j,m}/\left(\sum_{m=1}^{m=j_r} CF_{j,m}\right)$$

For the sake of illustration, the generation of interpretation rules for the −COOH substructure ($j = 2$, Table 2) is presented. According to above definitions and the correlation table given in Figure 3c:

$$j = 2 \qquad j_r = 2 \qquad j_{1\max} = 7 \qquad j_{2\max} = 2$$

$$CF_{2,1} = (60.29 + 59.00 + 51.02 + 51.00 + 49.22 + 46.82 + 23.98)/7 = 48.76$$

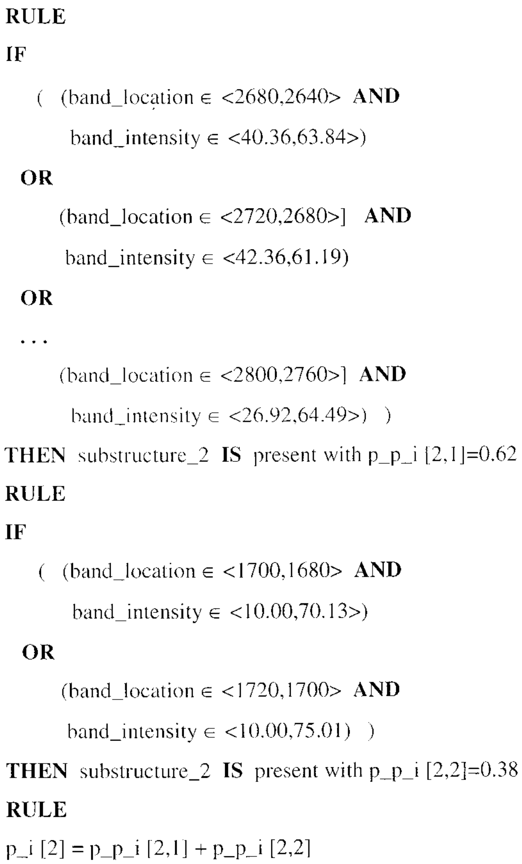
$$CF_{2,2} = (39.92 + 19.69)/2 = 29.81$$

$$\sum_{m=1}^{m=2} CF_{2,m} = CF_{2,1} + CF_{2,2} = 78.57$$

$$\text{partial\_probability\_index}[2,1] = 48.76/78.57 = 0.62$$

$$\text{partial\_probability\_index}[2,2] = 29.81/78.57 = 0.38$$

The substructure −COOH is recognized by two interpretation rules plus one rule for calculation of the probability index.



```
RULE
IF
   (  (band_location ∈ <2680,2640>  AND
        band_intensity ∈ <40.36,63.84>)
  OR
      (band_location ∈ <2720,2680>]  AND
       band_intensity ∈ <42.36,61.19)
  OR
  . . .
      (band_location ∈ <2800,2760>]  AND
        band_intensity ∈ <26.92,64.49>)  )
THEN  substructure_2  IS  present with p_p_i [2,1]=0.62
RULE
IF
   (  (band_location ∈ <1700,1680>  AND
        band_intensity ∈ <10.00,70.13>)
  OR
      (band_location ∈ <1720,1700>  AND
       band_intensity ∈ <10.00,75.01)  )
THEN  substructure_2  IS  present with p_p_i [2,2]=0.38
RULE
p_i [2] = p_p_i [2,1] + p_p_i [2,2]
```

This type of rules, related to spectral correlation, is then augmented by a set of rules required to control the inference engine (i.e., zeroing parameters, reading of input data, organizing cycles of the inference process by forward chaining, etc.) and the output of the obtained results. The complete knowledge base thus generated is ready to be used by the inference engine, a module SEK of the SCANKEE system.[58]

3. APPLICATION FOR SPECTRAL INTERPRETATION

During the consultation, the user enters spectral parameters, band locations, and intensities of the spectrum of an unknown compound. Alternatively, the spectral parameters can be extracted from the experimental spectra by dedicated programs specially developed at the Department of Computer Chemistry, Rzeszów (DCC). The inference engine, accessing the knowledge base, generates the list of substructures that would possibly constitute building parts of the investigated substructures, together with their calculated probability index. Obviously, the recognized substructures belong to the set of fragments for which the learning process was carried out. The scheme of the process can be presented as follows:

$$OC = \{(co_1, io_1), (co_2, io_2), ..., (co_m, io_m)\}$$
$$\downarrow$$
$$O = \{f_1, f_j, ..., f_k\}$$

The results of substructure identification obtained for compounds selected from the learning set enable the evaluation of the recognition process used to generate the rule knowledge base, but do not permit assessment of performance in generalizing to other compounds. As described above, the restriction of the *CF* vector to LEF correlation vector depends on the chosen threshold value. Therefore several knowledge bases can be generated for the same set of fragments (Table 2) but for different threshold values of the *CF* factor. Fixing a low threshold value increases the number of spectral features retained in the LEF correlation vector, leading to a greater number of fragments being incorrectly recognized as present. On the other hand, a too high threshold value leads to a decrease of correct positive answers. To optimize the threshold value that gives the best result, it is necessary to evaluate the overall performance of the resulting knowledge base using a collection of examples which were not included in the learning set. As the process of generation of a new knowledge base is automatic in the system, the algorithm to improve the existing rules is not yet implemented; thus, determination of the optimal threshold value has been conducted manually. A value ranging between 0.18 and 0.40, depending on the considered substructure, has been found as leading to a good compromise between the two drawbacks mentioned above (see Table 2).

4. RESULTS AND DISCUSSION

The main goal of this work was to evaluate the performance of the presented knowledge acquisition methodology and compare it with results obtained by other methods of substructure recognition. As was already pointed out by Soltzberg et al.[59] and will be recalled below, such comparison is rather difficult if the results are not obtained with the same collection of examples. We decided to compare the results obtained by the present approach to those obtained using hierarchical feed forward multilayer ANN[41] developed by one of us. To achieve a more reliable comparison, the same collection of IR spectra for learning and testing sets were used. Comparison with the performance of highly optimized flat feed forward multilayer ANN developed by Klawun[40] will also be presented, although the collection of compounds and spectra and experimental conditions used in this study are not the same. Although the knowledge acquisition method and the multilayer ANN were developed for a larger number of molecular fragments (222 and 33, respectively), the comparison will focus only on five substructures considered as characteristic, namely: hydroxylic, carboxylic, amino, benzenic, and ethylenic.

The computer database was prepared by building two independent sets of 1000 IR spectra each from the Sadtler Bio-Rad collection.[60] Only well-defined organic compounds were included in the two sets by random selection with the constraints that the distribution of the studied fragments does not differ significantly in the two sets. For each chemical compound the following information is stored: the structure of the compound coded as a connectivity matrix; the discrete curve of IR spectrum (resolution = 4 cm^{-1}), registered in absorption; the absorption band parameters (band location, band intensity) calculated using software elaborated in DCC.

One set was used as a *training set*, to generate the correlation tables for selected substructures (Table 2) which were in the rule knowledge base. The generated optimal knowledge base was evaluated in the process of prediction, using the second set as *testing set*.

Performance Evaluation. Although the problem of defining significant performance indices for binary classifiers was investigated about 25 years ago, one must note that the criteria, namely, the "information gain $I(A,B)$" proposed by Rotter and Varmuza[61] and the "figure of merit M" defined by Soltzberg et al.,[59] have not been widely used by investigators, at least in infrared spectroscopy.

Performance comparison of different methods is usually difficult because authors do not use the same indices of performance and it is not always possible to compute common indices on the basis of published results. Furthermore, the base of examples used for testing is not the same both in size and in composition; sometime even the experimental conditions are not the same. In particular, the percentage of correct responses which is commonly used as performance indices in the literature is highly dependent on the composition of the test base and can lead to overly optimistic results in the case of substructures poorly or highly represented in the test base used.

In the following we will use the same notation that we used in our previous work[41,42,62] and give the correspondence [within square brackets] with the notation used in ref 61.

For each substructure sought, let us note:

Nt: the total number of examples [N_{total}]

Np: (*N*umber of *p*resent) the number of compounds having the substructure [N]

Na: (*N*umber of *a*bsent) the number of compounds not having the substructure [$N_{\mathrm{total}} - N$]

cP: (*c*lassified as *P*resent) the number of compounds classified as having the considered substructure [N^{pred}]

cA: (*c*lassified as *A*bsent) the number of compounds classified as having not the considered substructure [$N_{\mathrm{total}} - N^{\mathrm{pred}}$]

PcP: (*P*resent *c*lassified as *P*resent) the number of compounds having the considered substructure and correctly identified as present [N^{corr}]

PcA: (*P*resent *c*lassified as *A*bsent) the number of compounds having the considered substructure and incorrectly identified as absent [$N - N^{\mathrm{corr}}$]

AcA: (*A*bsent *c*lassified as *A*bsent) the number of compounds without the considered substructure and correctly identified as absent [$N_{\mathrm{total}} - N - N^{\mathrm{pred}} + N^{\mathrm{corr}}$]

AcP: (*A*bsent *c*lassified as *P*resent) the number of compounds without the considered substructure and incorrectly identified as present [$N^{\mathrm{pred}} - N^{\mathrm{corr}}$]

On the basis of these quantities, various indices can be computed including the more frequently used GQ (*G*lobal *Q*uality), which is defined as the ratio of correct responses to the total number of examples.

$$GQ = (PcP + AcA)/Nt$$

$$GQ = [(N_{\mathrm{total}} + 2N^{\mathrm{corr}} - N - N^{\mathrm{pred}})/N_{\mathrm{total}}]$$

In fact this index suffers from the defect of being dependent on the balance of the test set. Therefore, in addition to this usual index, we have used the following indices which are more representative.

Pf: (*P*resent *f*ound) the fraction of compounds containing the considered substructure which was found correctly as present; $Pf = PcP/Np$

Af: (*A*bsent *f*ound) the fraction of compounds without the considered substructure and found correctly as absent; $Af = AcA/Na$

Table 3. Comparison of Performance Obtained by the Rule Knowledge Base and ANN

	Pf	Af	GQ	I(A/B)	M	EQr
			(a) This Work			
hydroxylic	0.667	0.789	0.745	0.148	0.157	0.446
carboxylic	0.950	0.850	0.887	0.501	0.527	0.594
amino	0.857	0.694	0.753	0.218	0.231	0.462
benzenic	0.907	0.877	0.888	0.468	0.497	0.756
ethylenic	0.278	1.000	0.849	0.141	0.190	0.544
			(b) Hierarchical ANN, Ref 41			
hydroxylic	0.833	0.933	0.897	0.466	0.495	0.776
carboxylic	0.879	0.931	0.912	0.525	0.552	0.812
amino	0.827	0.920	0.887	0.436	0.463	0.754
benzenic	0.864	0.790	0.828	0.339	0.339	0.655
ethylenic	0.426	0.936	0.829	0.106	0.144	0.484
			(c) Optimized ANN, Ref 40			
hydroxylic	0.967	0.990	0.980	0.846	0.858	0.950
carboxylic	1.000	0.969	0.972	0.335	0.780	0.823
amino[a]	0.886	0.980	0.972	0.283	0.658	0.824
benzenic	0.978	0.907	0.940	0.684	0.692	0.879
ethylenic	0.836	0.901	0.892	0.228	0.400	0.536

[a] Values for primary amine.

I(*A*,*B*): (*I*nformation gain) defined by Rotter and Varmuza[61] as the difference between *H*(*A*) the a priori uncertainty regarding class membership and *H*(*A*|*B*) the residual uncertainty after application of the classification method; $I(A,B) = H(A) - H(A|B)$

M: (figure of *M*erit) defined by Soltzberg et al.[59] as the ratio between *I*(*A*,*B*) and the maximum information gain which is itself equal to *H*(*A*); $M = I(A,B)/H(A)$

EQr: (*E*xtrastatistical *Q*uality of *r*esponses),[62] which is a measure of the relative improvement of the global quality over that which would have been expected by chance given the statistical distribution of the population.

$$EQr = (GQ - \mathrm{St})/(1 - \mathrm{St}) \quad \text{with}$$
$$St = 1 - 2Pi + 2Pi^2 \text{ and } Pi = Np/Nt$$

Pf and *Af* are significant because they are directly related to the reliability of the responses. *I*(*A*,*B*) has a firm theoretical basis. It measures the amount by which the classifier reduces the uncertainty regarding class membership. It is equal to zero when the classifier adds no information and its upper value is equal to the initial uncertainty, which depends on the composition of the test set.[59] *M* is the information gain relative to the maximum information gain imposed by the composition of the test set. As defined above, *EQr* is a measure of the relative improvement of the global quality over that which would have been expected by chance given the statistical distribution of the population.

The use of the last two indices *M* and *EQr* is recommended because they are not dependent on the composition of the test set used to compute them.

Results Comparison. The indices obtained by testing the generated knowledge base for the five considered substructures are shown in Table 3a. The best results of substructure recognition are obtained for carboxylic and benzene groups, as shown by the high values of their performance indices. The best performance indices of carboxylic and benzene groups allowed us to think that correlations generated for these substructures are highly significant.

The fairly high value of *EQr* obtained for the ethylenic substructure indicates that using the constructed rules improves the reliability of conclusions reached by the expert system compared to conclusions that would be obtained using only the statistical distribution of this substructures. However, a very low value of *Pf* factor indicates that the presence of the ethylenic group was found in a small number (27.8%) of compounds containing this molecular fragment. In fact, this substructure is poorly represented in the learning and testing databases (20.9%); therefore it is more difficult to construct rules to recognize the presence of this molecular fragment than to conclude to its absence. This case serve us to illustrate the above-mentioned limitation of the global quality index *GQ*. For the ethylenic group, $GQ = 0.849$, a quite good value (the system is right in 85% of the cases) hiding the poor performance at recognizing the presence of the substructure by the high performance at recognizing its absence.

The low *M* and *EQr* values of hydroxylic and amino groups seem to originate from the diffused definition of these substructures; i.e., their surroundings (e.g., aromatic ring, alkene bond, alkane groups) were not specified. Therefore, the generated correlations do not include vibrations of bonds linking the substructure with the remaining part of the molecule. This observation was already reported in the application of models based on similarity from Kohonen maps.[42] It was concluded from this study that the variability of spectra within a particular class, and the existence of subclasses such as alcohol, phenol, carboxylic acid, and enol within the general hydroxylic class, impose some limits on the models which can be constructed automatically from analysis of the spectral database. This applies to the clustering method used for projecting spectra on Kohonen maps, as well as to the construction of correlation tables presented and resulting rules.

For the sake of comparison, results obtained by application of layered neural network are also presented. Table 3b shows the results obtained by one of us using a hierarchical feed forward layered ANN,[41] while Table 3c shows results obtained by Klawun et al. using a highly optimized flat layered ANN.[40]

Drawing a comparison between these results, it can be stated that the global indices for hydroxylic, carboxylic, and amino groups obtained in the present work are lower than those resulting from the application of neural networks. On the other hand, for benzenic and ethylenic substructures the method of the rule generation gives slightly better results. Closer comparison shows that sometimes one approach is better than the other for specific indices. For example, for carboxylic substructure *Pf* is greater by rule generation than by hierarchical ANN, while the opposite is true for *Af*.

Comparison with flat optimized ANNs requires some caution, since the learning and testing sets are not the same both in size and in composition; also the spectra in this former work have been recorded in the gas phase, while spectra in the Sadtler collection we used are for liquid and KBr pellets. However, overall the results of Klawun et al. are better but the architecture of these flat networks cannot be modified to eventually incorporate other substructures, a limitation which does not apply to hierarchical ANN or to rule-based expert systems.

5. CONCLUSIONS

A new method for automatic construction of rules to predict the presence or absence of substructures from infrared spectra has been developed and evaluated. Strict comparison with the results obtained by hierarchical artificial neural networks, trained with the same set of examples, shows that ANN achieves better performance for some substructures while the opposite is true for others. To improve the performance of rule-based expert systems, it seems necessary to take into account the shape of the bands and not only their frequency and intensity. In conclusion, choosing the most favorable method of computer-aided substructure identification, with regard to a particular molecular fragment, should be based on detailed evaluations using significant indices of performances. This calls for a recommendation of setting a public database of spectra/structure couples that would allow the researchers in the domain to evaluate their approaches and compare to others using the same collection of examples.

Even if it is not reasonable to expect that any method of spectral interpretation may lead to 100% reliable results, its application does produce significant information gain. Therefore, the actual challenge of structure elucidation lies in the handling of uncertain information during the process of structure generation. The initial ambitious goal, stated in the 1970s, to develop fully automatic computerized systems of structure elucidation has been revised to design of interactive computer-assisted tools to help spectroscopists in their task and increase their productivity.

REFERENCES AND NOTES

(1) Warr, W. A. Computer-Assisted Structure Elucidation. 1. Library Search and Spectral Data Collections *Anal. Chem.* **1993**, *65*, 1045A−1050A.

(2) Warr, W. A. Computer-Assisted Structure Elucidation. 2. Indirect database Approaches and Established Systems *Anal. Chem.* **1993**, *65*, 1087A−1095A.

(3) Anderson, D. H.; Cover, G. L. Computer Search System for Retrieval of Infrared Data *Anal. Chem.* **1967**, *39*, 1288−1293.

(4) Erley, D. S. Quantitative Evaluation of Several Infrared Searching Systems. *Appl. Spectrosc.* **1971**, *25*, 200−202.

(5) Penski, E. C.; Padowski, A.; Bouck, J. B. Computer Storage and Search System for Infrared Spectra Including Peak Width and Intensity *Anal. Chem.* **1974**, *46*, 955−957.

(6) Clerc, J. T.; Zupan, J. Computer-Based Systems for the Retrieval of Infrared Spectral Data. *Pure. Appl. Chem.* **1977**, *49*, 1827.

(7) Azarraga, L. V.; Williams, R. R.; DeHaseth, J. A. Fourier Encoded Data Searching of Infrared Spectra (FEDS/IRS). *Appl. Spectrosc.* **1981**, *35*, 466−469.

(8) Woodruff, H. B.; Munk, M. E. A Computerised Infrared Spectral Interpreter as a Tool in Structure Elucidation of Natural Products. *J. Org. Chem.* **1977**, *42*, 1761−1767.

(9) Woodruff, H. B.; Smith, G. M. Computer Program for the Analysis of Infrared Spectra. *Anal. Chem.* **1980**, *52*, 2321−2327.

(10) Tomellini, S. A.; Hartwick, R. A.; Woodruff, H. B. Automatic Tracing and Presentation of Interpretation Rules Used by PAIRS: Program for the Analysis of IR Spectra. *Appl. Spectrosc.* **1985**, *39*, 331−333.

(11) Tomellini, S. A.; Wythoff, B. J.; Woodruff H. B. Developing Knowledge Base Systems. A Learning Process. In *Expert Systems Applications in Chemistry*; Hohne, B. A., Pierce T. H., Eds.; American Chemical Society: Washington, DC, 1989.

(12) Dębska, B.; Duliban, J.; Guzowska-Świder, B.; Hippe Z. Computer-Aided Structural Analysis of Organic Compounds by an Artificial Intelligence System. *Anal. Chim. Acta* **1981**, *133*, 303−318.

(13) Gribov, L. A.; Dementyev, V. A.; Elyashberg, M. E.; Yakapov, E. Z. Automation of Spectrochemical Investigations. *J. Mol. Struct.* **1974**, *22*, 161−172.

(14) Gribov, L. A.; Elyashberg, M. E.; Raikhshtat, M. M. A New Approach to the Determination of Molecular Spatial Structure Based on the Use of Spectra and Computers. *J. Mol. Struct.* **1979**, *53*, 81−96.

(15) Funatsu, K.; Sasaki, S. The Automated Structure Elucidation System CHEMICS. *Chem. Inf., Proc. Int. Conf.* **1989**, 271−281.

(16) Funatsu, K.; Sasaki, S. Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.

(17) Huixiao, H.; Xinquan, X. ESSESA: An Expert System for Elucidation of Structures from Spectra 1. Knowledge Base of Infrared Spectra and Analysis and Interpretation Programs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203−210.

(18) Andrev, G. N.; Argirov, O. K. Implementation of Human Expert Heuristics in Computer Supported Infrared Spectra Interpretation. *J. Mol. Struct.* **1995**, *347*, 439−448.

(19) Andrev, G. N.; Argirov, O. K. EXPIRS, an Expert System for Generation of Alternative Sets of Substructures, Derived by Infrared Spectra Interpretation. *Anal. Chim. Acta* **1996**, *321*, 105−111.

(20) Elyashberg, M. E.; Martirosian, E. R.; Karasev, Yu. Z.; Thiele, H.; Somberg, H. X-PERT: a User-Friendly Expert System for Molecular Structure Elucidation by Spectral Methods. *Anal. Chim. Acta* **1997**, *337*, 265−286.

(21) Seil, J.; Koehler, I.; von der Lieth, C. V. Interpretation of Infrared Spectra Based on Statistical Approaches. *Anal. Chim. Acta* **1986**, *188*, 219−227.

(22) Luinge, H. J.; Kleywegt, G. J.; van't Klooster, H. A.; van der Mass, J. H. Artifficial Intelligence Used for the Interpretation of Combined Spectral Data. 3. Automated Generation of Interpretation Rules for Infrared Spectral Data. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 95−99.

(23) Luinge, H. J.; van der Mass, J. H. Artifficial Intelligence for the Interpretation of Combined Spectral Data. Design and Development of the Spectrum Interpreter. *Anal. Chim. Acta* **1989**, *223*, 135−147.

(24) Ehrentreich, F.; Dietze, U.; Meyer, U.; Schulz, H.; Klötzer, H. M.; Abbas, S.; Otto, M. IR-Spectroscopy—Computer Program for the Interpretation of Infrared Spectra. *Fresenius J. Anal. Chem.* **1996**, *68*, 829−832.

(25) Ehrentreich, F. Representation of Extended Infrared Spectrum-Structure-Correlations Based on Fuzzy Theory. *Fresenius J. Anal. Chem.* **1997**, *357*, 527−533.

(26) Jamroz, M.; Latek, Z. The Algorithm of Substructure Recognition in Organic Compounds by the Artificial Intelligence Method on the Basis of IR Spectra. *J. Mol. Struct.* **1984**, *115*, 277−280.

(27) Trulson, M. O.; Munk, M. E. Table Driven Procedure for Infrared Spectrum Interpretation. *Anal. Chem.* **1983**, *55*, 2137−2142.

(28) Farkas, M.; Markos, J.; Szepesvary, P.; Barlta, I.; Szalontai, G.; Simon, Z. A Computer-Aided System for Organic Functional Group Determinations. *Anal. Chim. Acta* **1981**, *133*, 19−29.

(29) Passlack, M.; Bremser, W. In *Computer-Supported Spectroscopic Databases*; Zupan, J., Ed.; Ellis Horwood: Chichester, England, 1986.

(30) Affolter, C.; Baumann, K.; Clerc, J. T.; Schriber, H.; Pretsch, E. Automatic Interpretation of Infrared Spectra. *Mikrochim. Acta* [*Suppl.*] **1997**, *14*, 143−147.

(31) Roeges, N. P. G. *A guide to the Complete Interpretation of Infrared Spectro of Organic Structures*; Wiley: Chichester, 1994.

(32) Liddel, R. W.; Jurs, P. C. Interpretation of Infrared Spectra Using Pattern Recognition Techniques. *Anal. Chem.* **1974**, *46*, 2126−2130.

(33) Woodruff, H. B.; Ritter, G. L.; Lowry, S. R.; Isenhour, T. L. Pattern Recognition Methods for the Classification of Binary Infrared Spectral Data. *Appl. Spectrosc.* **1976**, *30*, 213−216.

(34) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta* (*Wien*) **1990**, *1*, 131−135.

(35) Munk, M. E.; Madison, M. S.; Robb, E. W. Neural Network Models for Infrared Spectrum Interpretation. *Mikrochim. Acta* (*Wien*) **1991**, *2*, 505−514.

(36) Meyer, M.; Weigelt, T. Interpretation of Infrared Spectra by Artificial Neural Networks. *Anal. Chim. Acta* **1992**, *265*, 183−190.

(37) Fessenden, R. J.; Györgyi, L. Identifying Functional Groups in IR Spectra Using an Artificial Neural Network. *J. Chem. Soc., Perkin Trans.* **1991**, *2*, 1755−1762.

(38) Ricard, D.; Cachet, C.; Cabrol-Bass, D. Neural Network Approach to Structural Feature Recognition from Infrared Spectra. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 202−210.

(39) Novic, M.; Zupan, J. Investigation of Infrared Spectra-Structure Correlation Using Kohonen and Counterpropagation Neural Network. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 454−466.

(40) Klawun, Ch.; Wilkins, Ch. L. Optimization of Functional Group Prediction from Infrared Spectra Using Neural Network. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 69−81.

(41) Cleva, C.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. Advantages of a hierarchical system of neural-networks for the interpretation of infrared spectra in structure determination. *Anal. Chim. Acta* **1997**, *348*, 255−265.

(42) Cleva, C.; Cachet, C.; Cabrol-Bass, D. Clustering of Infrared Spectra with Kohonen Networks. *Analysis* **1999**, *27*, 81−90.

(43) Gasteiger, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. Finding the 3D Structure of a Molecule in its IR Spectrum. *Fresenius J. Anal. Chem.* **1997**, *359*, 50−55.

(44) Schuur, J.; Gasteiger, J. Infrared Spectra Simulation of Substituted Benzene Derivatives on the Basis of a 3D Structure Representation. *Anal. Chem.* **1997**, *69*, 2398−2405.

(45) Zytkow, J. N.; Zembowicz, R. Database Exploration in Search of Regularities. *J. Intell. Inf. Syst.* **1993**, *2*, 39−81.

(46) Ziarko, W.; Shan, N. Database Mining Using Rough Sets *Proceedings Intelligent Information Systems IV, Augustów, Poland 1995*; Vol. 1, pp 74−86.

(47) Jones, R. N.; Bach, T. E.; Fuhrer, H.; Kartha, V. B.; Pitha, J.; Seshadri, K. S.; Venkataraghavan, R.; Young, R. P. Computer Programs for Absorption Spectrophotometry. *National Research Council of Canada, Ottawa 1971*, Vol. 11, pp 1−159.

(48) Jones, R. N.; Young, R. P. Additional Computer Programs for Absorption Spectrophotometry and Band Fitting. *National Research Council of Canada, Ottawa 1971*; Vol. 13, pp 1−129.

(49) Ehrentreich, F.; Nikolov, S. G.; Wolkenstein, M.; Hutter, H. The Wavelet Transform: a New Preprocessing Method for Peak Recognition of Infrared Spectra. *Mikrochim. Acta* **1998**, *128*, 241.

(50) Dębska, B.; Guzowska-Świder, B. Knowledge Discovery in an Infrared Database. *Comput. Chem.* **1997**, *21*, 51−59.

(51) Piottukh-Peletskii, V. N.; Derendyaev, B. G.; Bogdanova, T. F. Interpretation of IR Spectra Using Complete Sets of Fragment Compositions and IR Search System. 2. Determination of Microfragments Compositions of Organic Compounds. *J. Struct. Chem.* **1997**, *38*, 126−134.

(52) Piottukh-Peletskii, V. N.; Derendyaev, B. G.; Bogdanova, T. F. Interpretation of IR Spectra Using Complete Sets of Structure Fragment Compositions and IR Search System. 3. Analysis of Large Fragments. *J. Struct. Chem.* **1997**, *38*, 297−305.

(53) Piottukh-Peletskii, V. N.; Derendyaev, B. G.; Bogdanova, T. F. Complete Sets of Fragment Compositions of Structures for IR Spectrum Interpretation Using Database. 4. Forming the Most Probable Hypothesis about the Structure of the Unknown. *J. Struct. Chem.* **1997**, *38*, 657−664.

(54) Penchev, P. N.; Andreev, G. N.; Varmuza, K. Automatic Classification of Infrared Spectra Using a Set of Improved Expert-Based Features. *Anal. Chim. Acta* **1999**, *388*, 145−159.

(55) Varmuza, K.; Penchev, P. N.; Scsibrany, H. Large and Frequently Occurring Substructures in Organic Compounds Obtained by Library Search of Infrared Spectra. *Vib. Spectrosc.* **1999**, *19*, 407−412.

(56) Dementiev, V. A.; Timchenko, S. D. Chemical Structure and IR Spectra of Organic Compounds Connected by Statistical Methods. *J. Mol. Struct.* **1994**, *319*, 31−39.

(57) Michalski, R. S. Tutorial on Machine Learning, Data Mining and Knowledge Discovery. *Proceedings Intelligent Information Systems VII, Zakopane* 1997; Vol. 1, pp 1−87.

(58) Dębska, B.; Guzowska-Świder, B. Application of Knowledge Engineering Program Environment System SCANKEE for Recognition of Structural Units in the Molecule of an Organic Compound. *J. Mol. Struct.* **1995**, *348*, 473−476.

(59) Soltzberg, L. J.; Wilkins, C. L., Kaberline, S. L.; Fai Lam, T.; Brunner, T. R. Evaluation and Comparison of Pattern Classifiers for Chemical Applications. *J. Am. Chem. Soc.* **1976**, *98* (23), 7139−7144.

(60) Bio-Rad Laboratories Inc., Sadtler Division, 3316 Spring Garden Street, Philadelphia, PA, 19104, USA, Copyright 1980−1993. All rights reserved.

(61) Rotter, H.; Varmuza, K. Criteria for the evaluation of classifiers for the automatic interpretation of spectra (pattern recognition). *Org. Mass Spectrom.* **1975**, *10* (10), 874−884.

(62) Sbirrazzuoli, N.; Cachet, C.; Cabrol-Bass, D.; Forrest, T. P. Indices for the evaluation of neural networks performances as classifier: application to structural elucidation in infrared spectroscopy. *Neural Comput. Appl.* **1993**, *1*, 229−239.

CI990271X