

## New Computer Aided Methods for Revealing Structural Features of Unknown Compounds Using Low Resolution Mass Spectra

Konstantin S. Lebedev<sup>\*,†</sup> and Daniel Cabrol-Bass<sup>‡</sup>

Institute of Organic Chemistry, Siberian Branch of Russian Academy of Science, 9 Lavrentyev Avenue, Novosibirsk 630090, Russia, and Groupement de Recherche en Chimie Fine Organique (GRECFO), Laboratoire Représentation et Traitement de l'Information Chimique (LARTIC), Université de Nice Sophia-Antipolis (UNSA), Parc Valrose, 06108 Nice, Cedex 2, France

Received September 21, 1997

Two new computer methods designed to reveal structural features of unknown compounds by means of low resolution mass spectra are presented. Both methods use the results of a spectral similarity search in a mass spectral database. The first one proceeds by intersecting selected structures in order to find maximal common substructures, while the second proceeds by decomposing these structures to derive fragments following a model of primary fragmentation of organic molecules. Reliability of the revealed fragments is estimated by comparing an unknown compound's spectrum with the computed spectral images of each fragment. The usefulness and limitations of the two proposed methods are estimated by using a set of test examples. In many cases the two methods are complementary, whereas overall, the second looks more promising both for revealing large structural fragments and for generation of candidate structures, because the fragments revealed have only one or two free valences and rarely overlap one another.

### INTRODUCTION

In routine chemical analysis low resolution mass spectra are commonly used for identification of compounds which have been previously described. Computerized search systems greatly facilitate this task<sup>1–5</sup> by rapidly scanning the large volume of information available in mass spectral databases and retrieving compounds having spectra most similar to the one under investigation. As a rule, the final identification decision is made by an analyst on the basis of computed spectra similarity indexes and by visual comparison of an experimental spectrum with those retrieved through a database search procedure.

When the spectrum of a compound under study is missing from the reference database, the task becomes much more complex and is known as the "structure elucidation problem". Although intensive and continuous efforts have been devoted over the years to this problem by several research groups, it has not yet been completely solved. The trend in this field of research is to combine structural information derived from several spectroscopic data sets, leading to the so-called "multispectroscopic approach".<sup>6–8</sup> Also, and despite some recent claims,<sup>9</sup> the initial ambitious goal to construct fully automated systems is far from being reached and is being replaced by the more realistic prospect of developing decision-supported systems. In such a system, a human analyst makes use of various computer resources to increase his/her efficiency and productivity in the task of structure elucidation.<sup>10–13</sup> Besides the fact that decision-supported

systems are more likely to be accepted and used by analysts than fully automated systems, they offer the advantage of being open for continuous improvement and enhancement by addition of new modules.

In addition to resources that give access to various ways of searching spectroscopic databases and spectra–structure correlation tables, typically such systems incorporate specialized modules for (a) extraction of structural information from spectra of unknown compounds, (b) generation of candidate structures using some constraints (generally, but not exclusively, specified by good and bad lists of substructures) resulting from the previous step, and (c) verification and ranking of candidate structures to choose the most plausible one. Usually ranking is based on the comparison of simulated and observed spectra.

Contrary to other spectroscopic techniques, in particular <sup>13</sup>C NMR and IR, low resolution mass spectrometry has not been exploited to its full potential up to now, although this analytical method was the first to be used in the pioneer DENDRAL project.<sup>14</sup> This is due to the complexity of the relationship between a compound's structure and mass spectra. Although efficient classifiers based on numerical transformation of spectra and neural-networks<sup>15–18</sup> or other methods such as multivariate discriminate analysis have been reported,<sup>19</sup> revealed structural features are usually of limited size. Undoubtedly, application of these classifiers are helpful to reduce the numbers of candidate structures during the structure generation phase. But there is a pressing need for new methods to reveal structural features of larger size in order to take full advantage of low resolution spectra.

<sup>†</sup> Siberian Branch of Russian Academy of Science.

<sup>‡</sup> Université de Nice Sophia-Antipolis (UNSA).

Currently, three main approaches have been developed to tackle the problem of revealing structural features using spectra search results (SS-results). The first one analyzes structures from SS-results to identify fragments from a previously prepared list;<sup>20,21</sup> the second reveals fragments used to describe structures of compounds stored in a computer database,<sup>22</sup> and the third one constructs fragments as maximal common substructures by intersecting pairs of structures from SS-results.<sup>23–25</sup>

Despite some differences, all these approaches estimate the reliability of revealed substructures on the basis of the frequency of their appearance in structures of SS-results. A list of fragments ranked in accordance with this reliability criterion is then made available to the investigator. It is important to emphasize that the value of this criterion is properly the sole argument when deciding among suggested fragments, which may include incorrect ones. Literature data analysis and our own work showed that it is almost impossible to eliminate incorrect solutions using only data on frequencies of fragment appearance in SS-results, even if statistical peculiarities of databases are taken into account.<sup>20,22,26</sup>

Therefore, further development of computer methods for the analysis of low resolution mass spectra should invoke additional information at the stage involving the ranking and viewing of results. For this purpose, referring to spectral data is likely to be most advantageous. This is most evident in the case of NMR spectroscopy, thanks to its ability to directly assign spectral signals to distinct atoms in a molecule, as for example in ref 27.

A similar approach has proved to be efficient in the case of IR spectroscopy as well.<sup>8,28</sup> Along with a substructure common to several structures of SS-results an investigator obtains a set of absorption bands, or “spectral image”, which are common to spectra of corresponding compounds. Experiments in solving test problems have demonstrated that comparison of the spectra of the compound under study to the “spectral image” of revealed substructures, which incorporates knowledge of IR correlations, allows a less fallible decision than in cases where only data on the appearance frequency of these substructures in the SS-results are used.

In this paper we present and discuss two new methods of analysis SS-results to reveal large structural fragments of unknown compounds from low resolution mass spectra.

## METHODS

**Initial Spectral Search.** Both methods rely on retrieving from the mass spectral database compounds whose spectra are most similar to the spectrum of the unknown compound leading to the SS-result list. For this purpose, original search procedures have been developed. Similarity of compared spectra is estimated by means of the spectral match factor ( $\leq 100\%$ ), similar to MF 10 of McLafferty.<sup>29</sup>

The search procedure makes use of a set of parameters which allow for consideration different spectral data ( $m/z$  ions, masses of neutral losses, intensities of the corresponding peaks<sup>5</sup>) in the comparison and can apply various search methods<sup>26</sup> (forward, reverse, and combined). The procedure

parameters and search method are specified by investigator depending on the characteristics of the problem at hand; a set of default parameters and standard method is offered. As a result, the search procedure built the list of reference compounds (SS-results) ranked in decreasing order of their spectral match factor with the spectra of the unknown. For further analysis, mass spectra and structures of the  $n$ -first compounds from the complete SS-results list are retained. In our experiment a value of  $n \leq 50$  was used.

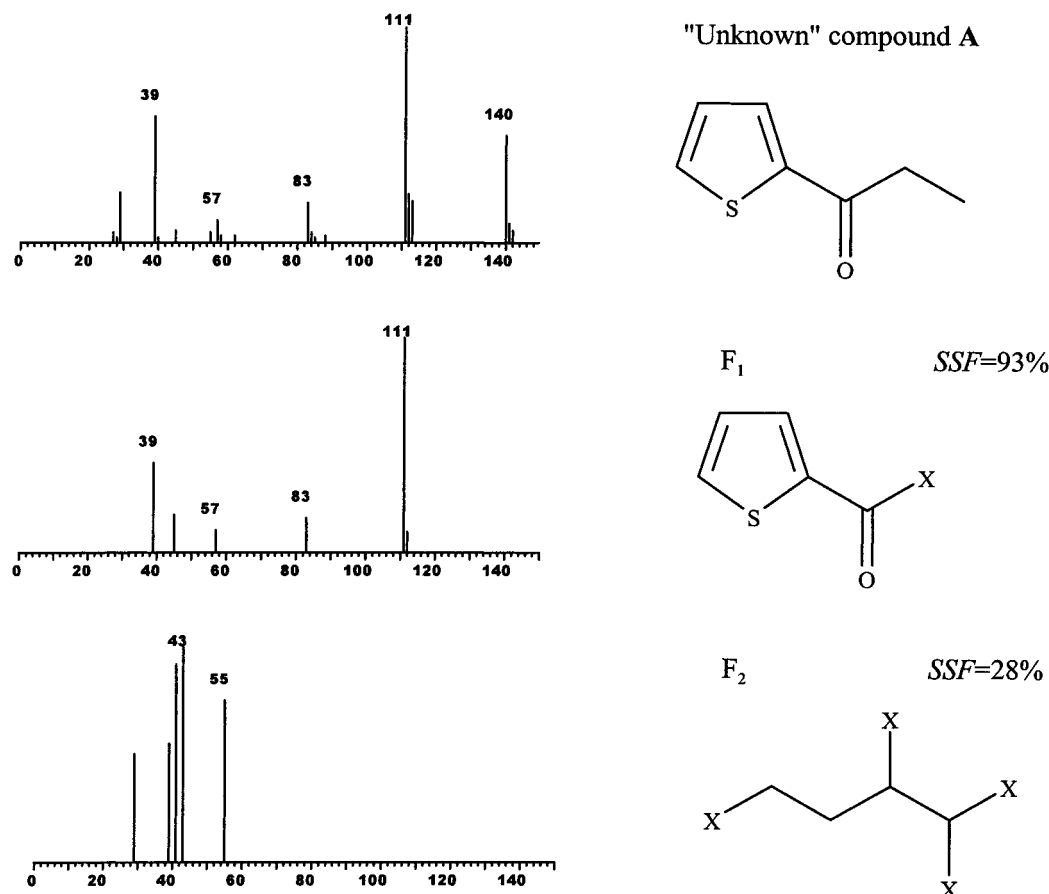
**Intersections of Structures of SS-Results.** The first method we have developed requires (a) determination of a maximum common substructure (MCSS) in the compound's structures from SS-results, (b) construction of a “spectral image” and (c) estimation of the reliability of the revealed fragments. In the following discussion this method will be called the MCSS-based method. Because general procedures for MCSS searches are well established, only a short description of its adaptation to the specific problem at hand is given here.

In the MCSS-based method, the first  $n$  structures of SS-results are intersected with each other to find common substructures with no less than  $p$  nodes (a total of  $n(n-1)/2$  intersections are examined). A simple algorithm based on an in-depth enumeration of all possible node-by-node matches is used for this purpose. However, the minimal number of nodes,  $p$ , is allowed to increase for each pair of structures as the algorithm continues to search for the MCSS. In our case this is required because the procedure is designed to find the largest substructures possible. Small values of  $p$  could result only in accidental identification of several small fragments, whereas all the maximal ones will appear with greater certainty.

As a result of the first step a list of nonisomorphic fragments is obtained. Next, a “spectral image” for each fragment is formed. A “spectral image” is made of a set of spectral peaks occurring most often in spectra that correspond to structures of SS-results containing a given fragment (the maximal number of peaks and the minimal frequency of peaks are parameters of the method). Reliability of derived fragments is estimated by comparing their spectral images with the spectrum of an unknown. We proceed from the assumption that the more closely a “spectral image” of a fragment matches a spectrum of an unknown, the higher is the reliability of the fragment, i.e., the greater the probability of its presence in an unknown structure. The following formula, similar to McLafferty's MF10, is applied to calculate the spectral similarity factor  $SSF$ , which is used to rank revealed fragments

$$SSF = \frac{\sum (x + r) \frac{\min(x, r)}{\max(x, r)}}{2 \sum r} \quad (1)$$

where  $x$  and  $r$  represent the logarithms of intensities of peaks in an experimental spectrum and a “spectral image”, respectively. The summation in the numerator applies to the peaks which coincide by the  $m/z$  value, while in the denominator it applies to all peaks constituting the fragment's “spectral image”.



**Figure 1.** Two fragments with their "spectral image" revealed by MCSS based method for "unknown" compound A. *Note:* Here and below, X in the structural fragments denotes a substituent other than hydrogen.

To illustrate this method, two substructures ( $F_1$ ,  $F_2$ ) and their corresponding "spectral images" derived from the analysis of SS-results obtained for "unknown" compound A are given in Figure 1. In this case both correct ( $F_1$ ) and incorrect ( $F_2$ ) fragments occur to the same extent (i.e., six times) in the first 45 structures constituting SS-results, whereas the spectral image of  $F_1$  is in better agreement ( $SSF = 93\%$ ) with spectrum A than that of  $F_2$  ( $SSF = 28\%$ ).

Moreover, the "spectral image" of  $F_1$  is not in contradiction with general considerations of mass spectrometry. Indeed, the peaks with  $m/z$  111 ( $C_5H_3SO$ ) and 83 ( $C_4H_3S$ ) can be easily explained by the fragmentation of the molecular ion  $m/z$  140 ( $C_7H_8SO$ ) connected with the elimination of  $C_2H_5$  and  $C_2H_5CO$  groups and peaks with  $m/z$  45 (CHS) and 39 ( $C_3H_3$ ) by the destruction of thiophenyl ring.

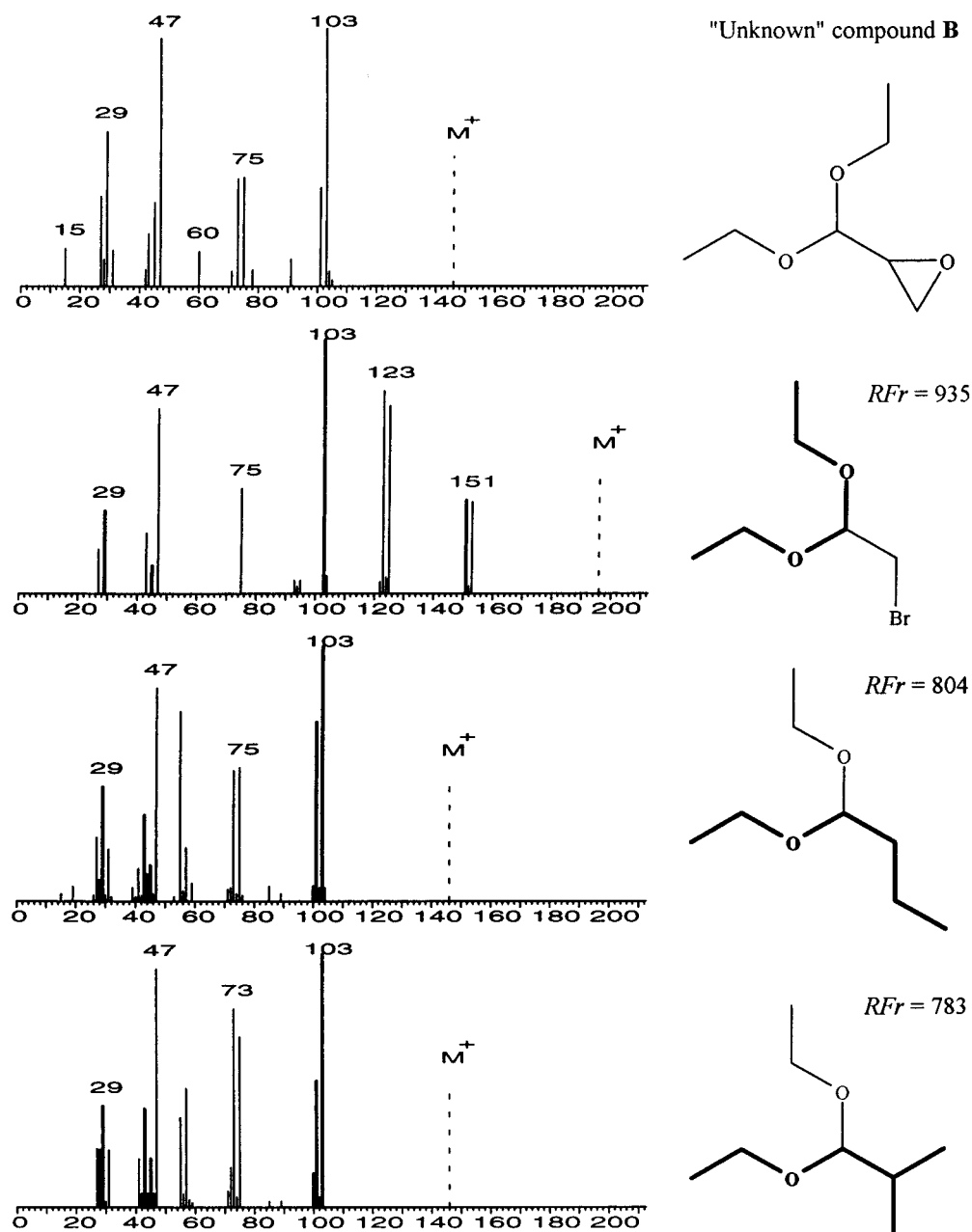
Experiments on a wealth of examples have shown us that, despite its formal nature, this method can lead to useful insights into the structure of unknown compounds, which could not easily be found by manual inspection of low resolution mass spectra.

Nevertheless, pitfalls were found when applying the MCSS-based method in a number of cases. First, this method requires at least two structures with a common fragment to be present in SS-results. Second, since the method finds only common substructures with a number of connected nodes greater than a given value, it can lead to incorrect fragments which might include correct ones of a somewhat smaller size. And last, analysis of spectra of compounds bearing a fragment under consideration can

reveal spectral features not directly attributable to a given fragment (see, for example, in Figure 1 the peak with  $m/z$  57 in the "spectral image" of  $F_1$ ), a situation which makes a correct decision difficult to make. To overcome the above disadvantages we have developed another method for the analysis of SS-results based on modeling the first stage of fragmentation of organic molecules under electron ionization.

**Decomposition of Structures of SS-Results.** Let us recall that the primary fragmentation stage (PFS) consists essentially of decomposing a molecular ion ( $M^{+\bullet}$ ) into two parts by cleaving one or several bonds to produce odd-electron ( $A^+$ ) and even-electron ( $B^{+\bullet}$ ) ions in accordance with the following schemes:  $M^{+\bullet} \rightarrow A^+ + R^\bullet$  and  $M^{+\bullet} \rightarrow B^{+\bullet} + m^\circ$ , where  $R^\bullet$  is a radical and  $m^\circ$  is a neutral molecule. These processes are both simple for modeling and informative for solving structural problems, covering practically all the variety of molecules decomposed by electron ionization.

The suggested method (hereinafter called the PFS-based method) for revealing fragments of unknown structure from SS-results includes the following operations. The PFS modeling procedure treats independently each structure and corresponding mass spectrum from SS-results. A structure is cleaved in two parts by removing one or two bonds in all possible ways (cf. work<sup>30</sup>). Double and "hanging" (i.e., leading to a terminal vertex) bonds usually remain intact, though cleavage is an option. One can also choose how to break bonds in aromatic rings: any bond, only bonds adjacent to a substituted vertex, or none.



**Figure 2.** First three structural fragments revealed by PFS based method for "unknown" compound B. Note: Fragments and peaks constituting their "spectral image" are drawn in thick lines on the complete structure and spectra of the compound retrieved from SS-results.

During the second stage, each resulting fragment passes a simple check of consistency with the spectrum of compounds from SS-results. Let  $m/z$  represent ion mass number,  $\Delta m$  the primary neutral mass loss ( $\Delta m = MM - m/z$ , where  $MM$  is the molecular mass of given compound), and  $M_F$  the mass of the fragment in question. Only those fragments for which  $|M_F - m/z| \leq 2$  or  $|M_F - \Delta m| \leq 2$  are retained for further analysis, subject to the condition that the corresponding peak intensities are above the threshold value (in our experiments that value was 10% of the most intense peak). Fragments for which masses differ from  $m/z$  or  $\Delta m$  by less than two amu were retained in the final list to take into consideration fragmentation processes involving the migration of one or two hydrogen atoms, such as the McLafferty rearrangement and the elimination of neutral molecules such as  $H_2$ ,  $H_2O$ , and  $HCl$ . Let us consider, for example, a

compound of  $MM = 100$  and containing a propyl (or isopropyl) group in its structure. If the corresponding spectrum contains peaks at  $m/z = 42, 43, 44, 56, 57$ , and  $58$ , then the two peaks  $m/z$  and  $\Delta m$  from the propyl group would account for all six peaks within an absolute difference of one amu. Thus, for each compound from the SS-results, a list of structural fragments, which have been assigned to its mass spectrum ion peaks and (or) "neutral losses" peaks, is found. In other words we obtain "spectral-structure" correlations, which are attributable to molecular ion primary fragmentation.

The next step aims to obtain more evidence for larger fragments. If a fragment contains some smaller ones with the same type of supporting data (i.e.,  $m/z$  or  $\Delta m$ ) then it is assigned all peaks accounting for its subfragments. One can apply various physical meanings to this assumption. For

example, parallel fragmentation processes could provide additional support. Thus a fragment can formally account for many peaks that can be related to both its own mass and masses of its inner subfragments. As for the MCSS-based method, this set of peaks is called the fragment's "spectral image" resulting from PFS model. Extending the previous example with additional peaks at  $m/z = 28$ , 29, and 30 which correspond to a ethyl group ( $M_F = 29$ ), one could relate all these three peaks to the propyl group. However, one could not do the same with an isopropyl group, as it does not contain an ethyl group.

In the final step, the revealed fragments are ranked in order to select the most plausible ones for the compound under study. Let  $St_x$  be the unknown structure,  $St_r$  a structure from SS-results, and  $MS_x$  and  $MS_r$  their respective mass spectra. We have based our definition of a reliability factor on the following logical considerations:

- if a spectral feature ( $m/z$ ,  $\Delta m$ ) of the fragment's "spectral image" is present in  $MS_x$ , then there is some probability that this fragment belongs to  $St_x$ .
- the more spectral features "assigned" to the revealed fragment that are found in  $MS_x$  and the better they agree in corresponding peaks intensities, the greater is the probability that the given fragment is a part of  $St_x$ .

To estimate the reliability of the revealed fragments, each fragment is processed twice: first with  $MS_r$  of the structure  $St_r$ , from which it originates and, second, with  $MS_x$  of the unknown structure  $St_x$ . Thus every fragment has two "spectral images" that contain parts of  $MS_r$  and  $MS_x$ . The two "spectral images" are then compared to compute a reliability factor ( $RFr$ ) using the following eq 2

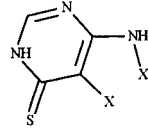
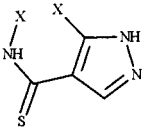
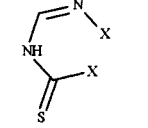
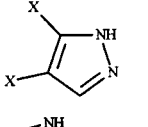
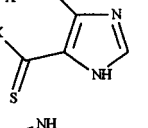
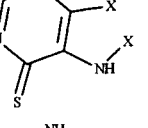
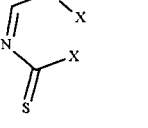
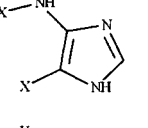
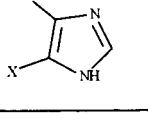
$$RFr = \frac{\sum(x' + r') \frac{\min(x', r')}{\max(x', r')}}{\sum x' + \sum r'} \cdot \sqrt{\frac{\sum x' + \sum r'}{\sum x + \sum r}} \quad (2)$$

where  $x$  and  $r$  represent logarithms of intensities of all peaks in compared spectra, while  $x'$  and  $r'$  represent logarithms only of those peaks belonging to "spectral image". This equation is similar to eq 1, with an additional term designed to favor fragments whose "spectral images" are made of many peaks. This factor was introduced to rank the largest possible fragments, which have spectral features most appropriate to the spectrum of the compound under study, at the top of the list.

Application of the PFS model to the SS-results yields a list of fragments with assigned numeric factors  $RFr$ . For each fragment in this list, the program displays the structure of the compound from the SS-results from which it has been derived and the corresponding mass spectrum. To facilitate the investigator's decision, the revealed fragment and the ions peaks and (or) "neutral losses" peaks assigned to the given fragment are highlighted.

As an example Figure 2 shows the first three structural fragments from the ranked list of fragments revealed by PFS based method applied to the SS-results obtained for the mass spectrum of "unknown" compound **B**. One can easily see that the correct fragment heads the list and that all the spectral features assigned to this fragment ( $m/z = 29$ , 45, 103, 151;

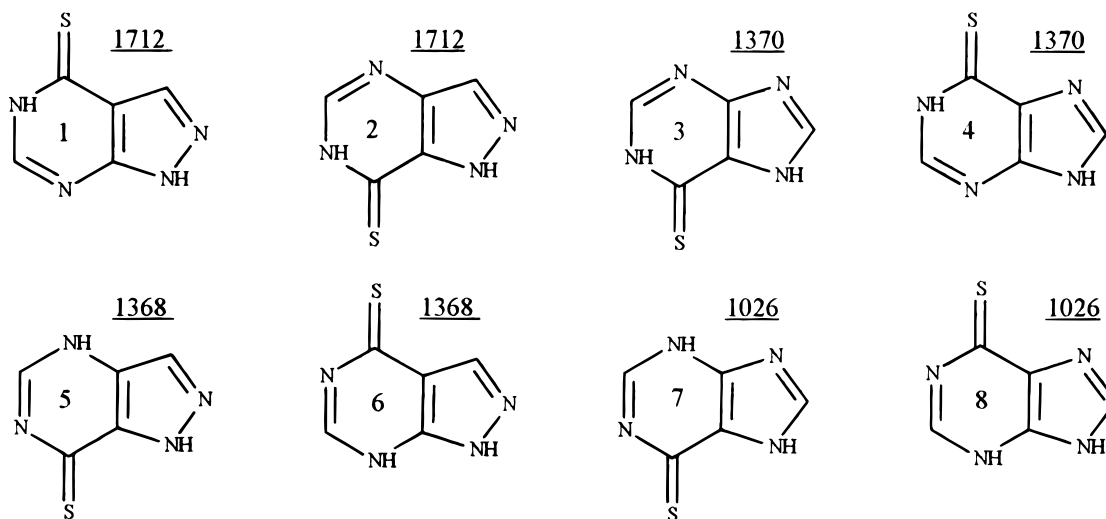
**Table 1.** First Nine Fragments Revealed for "Unknown" Compound C by PFS Based Method

	The structural fragments	Mass	El. Composition	<i>RFr</i>
F1		125	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> S	907
F2		125	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> S	907
F3		86	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub> S	871
F4		66	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub>	841
F5		125	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> S	595
F6		125	C <sub>4</sub> H <sub>3</sub> N <sub>3</sub> S	595
F7		86	C <sub>2</sub> H <sub>2</sub> N <sub>2</sub> S	527
F8		81	C <sub>3</sub> H <sub>3</sub> N <sub>3</sub>	516
F9		66	C <sub>3</sub> H <sub>2</sub> N <sub>2</sub>	499

the last corresponding to  $\Delta m = 45$ ) are present in the mass spectrum of compound **B**. This result is a strong argument in favor of the fragment suggested by the program. However, it should be noted that incorrect fragments that are also in good agreement with the analyzed mass spectrum are derived together with the correct ones.

## EXPERIMENTAL SECTION

**Implementation.** The software modules used to apply the methods described in this paper form separate applications of the ChemArt system.<sup>31</sup> All computer experiments were executed within the framework of this system. It both facilitates an independent upgrade of distinct procedures and the combination of modules without implicit transfer of intermediate data. Specialized modules of the ChemArt



**Figure 3.** Ranked list of generated structures for the "unknown" compound C. *Note:* numbers above structures are the reliability estimated value used for ranking.

system have been used in this study: for spectral similarity search, selection of structural analogues of unknown compounds, the handling of information about mass numbers of ions, masses of primary neutral losses and intensities of corresponding peaks; and the generation of structural formulas GENS.<sup>32</sup> All the modules were written in C and Pascal languages and run under Windows.

Because these methods are conceived as tools to be used by a specialist to support his or her decisions in the complex task of structure elucidation, the ergonomics of the user interface are very important. The ChemArt system takes full advantage of the Windows environment to construct a comprehensive user-friendly interface to the various modules constituting the system. Choices of numerous parameters controlling the operations of all modules, access to individual spectra and structures, handling of lists of SS-results, visual inspection of subspectra-substructure pairs, selection and rejection of retrieved substructures, and activation of the structure generator are under user control through menus and dialogue windows.

**Database and Initial Data.** The database of Novosibirsk Institute of Organic Chemistry (NIOC) containing approximately 50 000 mass spectra and corresponding structures of organic compounds was used as the reference database to carry out spectral searches. Assessment of the developed methods was conducted using mass spectra of (1) compounds synthesized at NIOC, (2) volatile fragrance compounds studied in UNSA, and (3) the database of LARTIC. When a compound under study was present in the reference database, it was temporarily removed during the assessment of the methods.

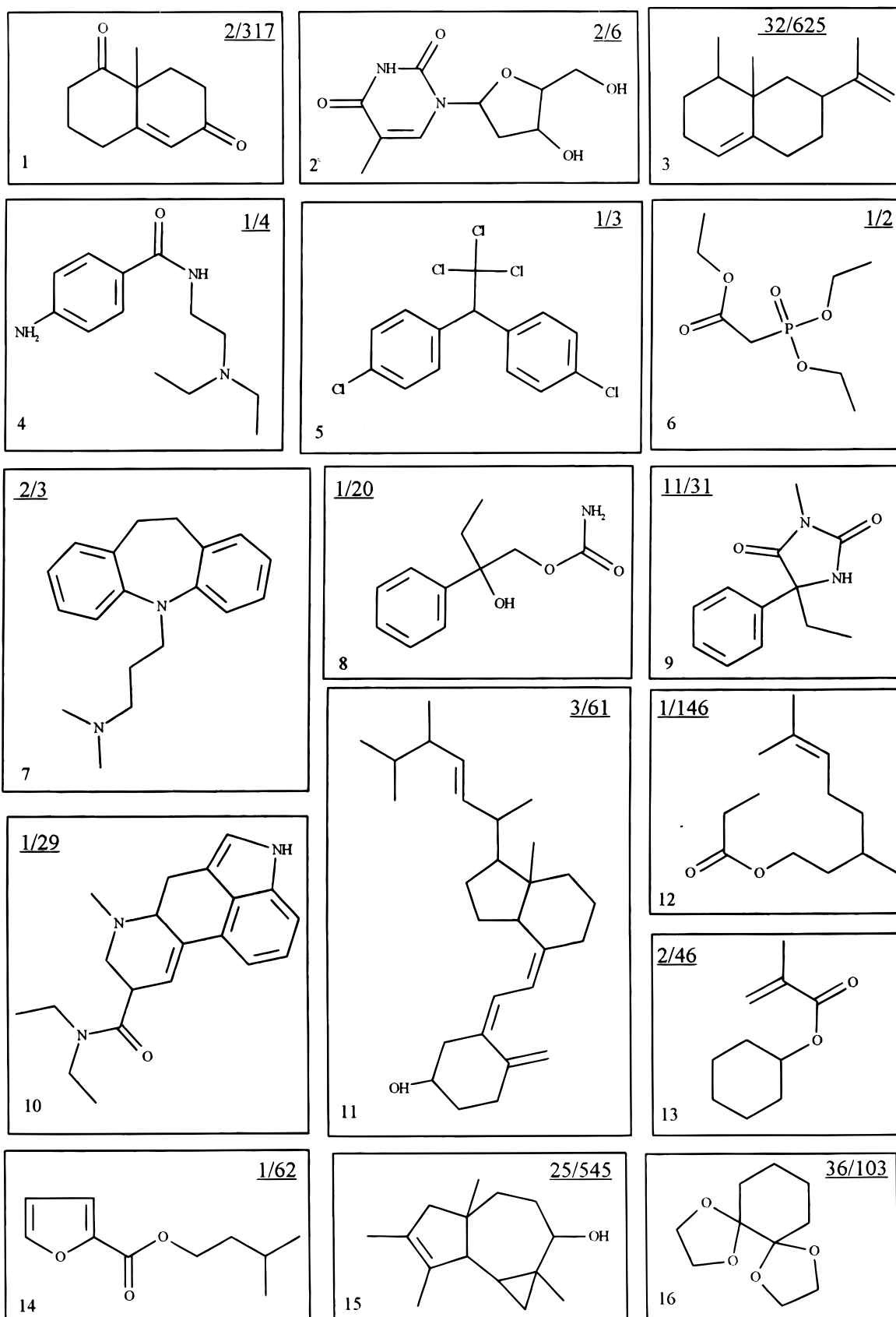
**Methods Efficiency Assessment.** Because the system does not operate in an automatic mode and requires active participation of a human expert both for choosing parameters and for analyzing obtained results, it is not possible to obtain significant statistical indices of performance. Furthermore the methods rely heavily on the presence in the reference database of compounds with structures similar to an unknown. Thus the composition of the set of examples used for testing would have a very strong influence on the

results. As far as we know, the problem of objective selection of a test compounds set is not yet resolved. Developing such a set would require the use of similarity index measures for clustering a reference spectra-structural database prior to the selection of a set of examples truly representative of the diversity of the available database. Even if this is done, the reliability of results would still be related to the reference database used. Obviously, structure elucidation of compounds which fall within a cluster of structures in a reference database is more likely to be successful than in the case of compounds which fall in "holes" between clusters.

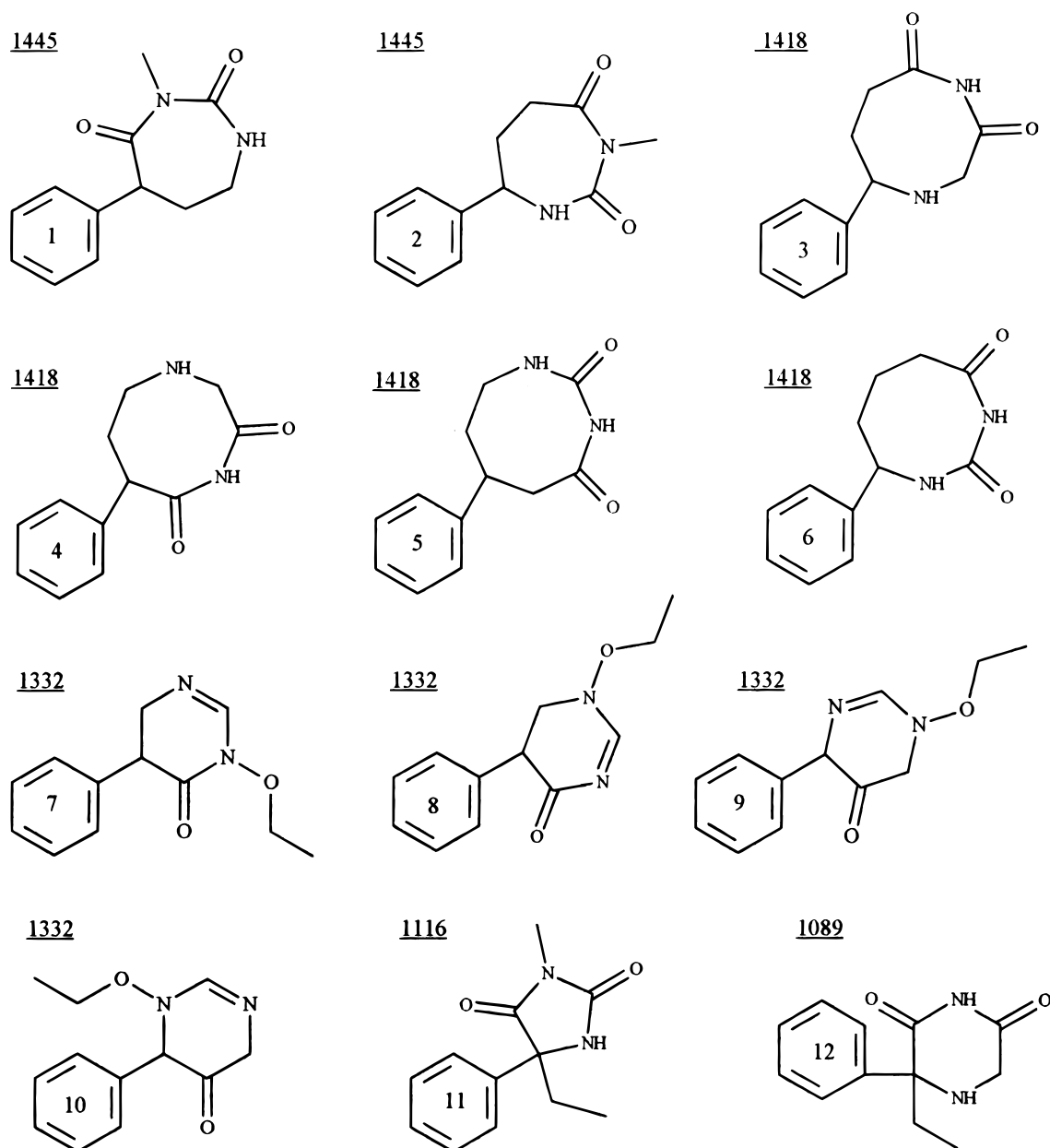
Therefore the results reported below should be considered an illustration of the potentials and limits of the proposed methods rather than results of statistical tests.

## RESULTS AND DISCUSSION

Extensive investigations carried out with a wealth of examples showed that it is nearly impossible to avoid the presence of false responses in a list of revealed fragments, even by changing the numerous parameters of the methods. Hence an investigator has to make the ultimate decision in selecting fragments when using both the MCSS and PFS-based methods. From the user's point of view, the difference lies in the fact that the "spectral images" built in applying the PFS-based method are better-substantiated and allow more straightforward manual interpretation. The PFS-based method also offers unquestionable benefits by its ability to lead to a solution in cases for which one compound very similar (in both the spectral and the structural sense) to an unknown is retrieved in the SS results. In such cases, rather large fragments describing almost entirely the structure of a compound under investigation are likely to be recognized. Moreover, in many cases, the PFS-based method allows the identification of nonoverlapping pairs of fragments, which might give a whole structure by simple combination. As an example Table 1 shows nine fragments revealed from the analysis of SS-results of the mass spectrum of "unknown"

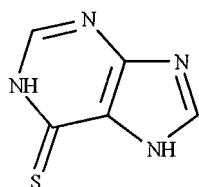


**Figure 4.** Structures of 16 “unknown” compounds which have been determined using molecular formula and mass spectrum. *Note:* Numerator = rank of the “unknown” structure in the list of generated structures; denominator = total number of generated structures on the basis of pairs fragments revealed by PFS method.



**Figure 5.** The first 12 structures from the ranked list obtained for structure elucidation of “unknown” compound N9 (see Figure 4).

compound **C** with molecular mass  $MM = 152$ , and molecular formula  $MF = C_5H_4N_4S$



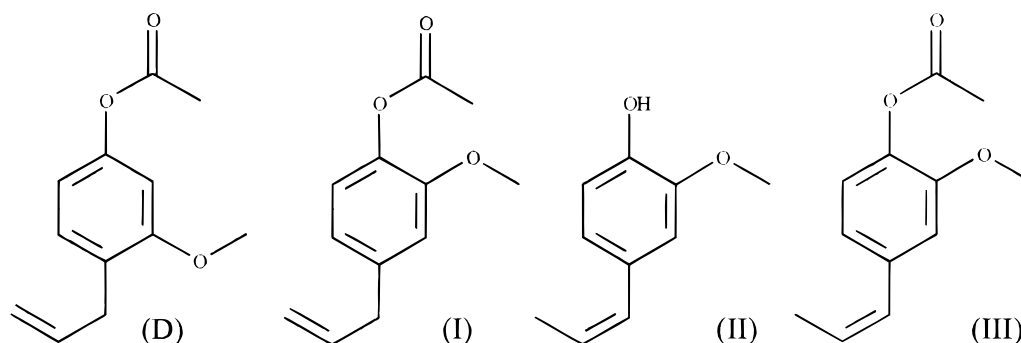
In this case four pairs of nonoverlapping fragments (F3 and F4, F3 and F9, F4 and F7, F7 and F9) compatible with the given  $MF$  can be found among the revealed fragments. These pairs provide a basis for the construction of eight molecular structures, including the “unknown”. It is possible to rank resulting structures by summing the reliability factors of each fragment constituting these complete structures (see Figure 3). According to this simple estimation, the

“unknown” structure takes the third place in the ranked list of the solutions. It worth noting that in this particular case it would be very difficult to discriminate between the proposed isomers, even by using other spectroscopic data.

As additional examples Figure 4 gives the structures of 16 “unknown” compounds successfully built among other candidate structures using the PFS based method. To construct automatically candidate structures for these test problems, only pairs of fragments whose total elemental composition deviates from the unknown’s molecular formula by no more than one heavy atom have been used as input to the structure generator GENS.<sup>32</sup> Fractions reported in this table represent the following: as numerator the position of the correct structure in the ranked list of generated structures and as denominator the total number of generated structures. Examination of these numbers shows that the correct solution does not always occupy the first place, and in the worst of



Chart 1



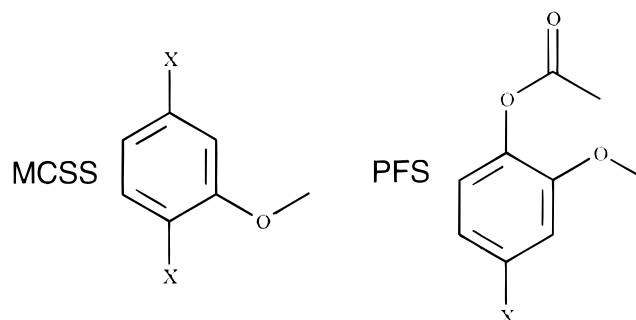
indicated examples (case 3), the correct structure is only at rank 36 in a list of 103 structures. However, considering the fact that these results are obtained using only mass spectra and molecular formula of compounds under study, they can be considered quite satisfactory, since low resolution mass spectrometry is not yet expected to provide an unequivocal answer for a structural task.

It should be kept in mind that this approach is proposed here as one possible component of a decision-supported structure elucidation system. Taking into consideration additional spectral data should help an analyst to avoid many false decisions and as a consequence would reduce drastically the number of candidate structures. To illustrate this point, we give in Figure 5 the first 12 structures from the ranked list of solutions, obtained by analysis of low resolution mass spectra of the "unknown" compound N9 (see Figure 4), the correct one being at rank 11. It is easy to see that the availability of  $^1\text{H}$  NMR and/or  $^{13}\text{C}$  NMR spectrum would allow an analyst to reject the first 10 candidate structures, as they do not contain the proper number of methyl groups (easily recognized using NMR data), and find the correct structure at the top of the ranked list.

However, it should be noted that the analysis of SS-results does not always reveal fragments which, when combined by pairs, allow the nearly complete reconstruction of the unknown structure. In such cases, the number of sets of fragments formed satisfy the molecular formula of the investigated compound and, as a consequence, the number of structures generated will grow quite rapidly. Therefore, taking into account additional structural information deduced from other approaches such as classifiers and/or from other spectroscopic data would help to rule out fragments from the list and limit the risk of combinatorial explosion. The large sizes of fragments revealed by our methods would facilitate this elimination procedure.

Comparison of the two methods using a set of examples showed that, as a whole, PFS is preferable to MCSS from the point of view of obtaining useful structural information about investigated compounds and in the subsequent use of revealed fragments at the generation stage, as they have only one or two free valences. However in some cases it turned out that the MCSS-based method gave better results. This might be expected in two situations: when application of the primary fragmentation model does not apply or when suitable structural analogues of a compound under study are absent from the reference database. As an example of such a situation, the first three structures I, II, and III from the SS-results of "unknown" compound **D** are shown in Chart 1.

Application of the MCSS-based method allowed the rather large fragment of compound **D**, to be revealed, while PFS gave a false result.



It is also worth mentioning that in the majority of our experiments the two proposed methods gave results which confirm and enhance each other.

## CONCLUSION

The methods we have developed for revealing structural fragments of unknown compounds from low resolution mass spectra still have some limitations and work is in progress for further improvement. In the case of the MCSS-based method it might appear useful to consider the statistical frequency of appearance of revealed fragments contained in the reference database, and the PFS-based method could benefit from the explicit use of empirical knowledge about the features of primary processes of fragmentation in some classes of organic compounds (cf. work<sup>33</sup>). Further development of our approaches is related to the combined use of fragments revealed by the two methods both for the generation of candidate structures and for their ranking.

## ACKNOWLEDGMENT

The authors thank Igor Stokov from NIOC for allowing the use of some software and for useful discussions about this work. The cooperation between NIOC and LARTIC was made possible by a NATO grant awarded to K.L. for his stay at Université de Nice Sophia-Antipolis. This work was supported in part by the Russian Basic Research Foundation (Grant 97-03-33514). Dr. Loretta Jones (University of Northern Colorado) is gratefully acknowledged for her help with the linguistic revision of the manuscript.

## REFERENCES AND NOTES

- (1) Zupan, J. *Computer-Supported Spectroscopic Databases*; Zupan, J., Ed.; Halsted: New York, 1986.
- (2) Warr, W. A. Spectral Databases. *Chemom. Intell. Lab. Syst.* **1991**, *10*, 279–292.
- (3) Martinsen, D. P.; Song, B.-H. Computer Applications in Mass Spectral Interpretation: A Recent Review. *Mass Spectrom. Review* **1985**, *4*, 461–490.
- (4) Domokos, L.; Henneberg, D.; Weimann, B. Computer-Aided Identification of Compounds by Comparison of Mass Spectra. *Anal. Chim. Acta* **1984**, *165*, 61–74.
- (5) Lebedev, K. S.; Kirschanskii, S. P.; Nekhoroshev, S. A.; Derendyaev, B. G. Computerized analysis of Mass Spectra for Chemical Structure Elucidation. Analytical Possibilities of the COMPAS-MS System. *Zh. Anal. Khim.* **1987**, *42*, 1320–1329 (Russ).
- (6) Bremser, W.; Fachinger, W.; Multidimensional Spectroscopy. *Magn. Reson. Chem.* **1985**, *23*, 1056–1071.
- (7) Gray, N. A. B. *Computer-Assisted Structure Elucidation*; Wiley: New York, 1986.
- (8) Lebedev, K. S.; Derendyaev, B. G. Computer Methods for the Solution of Structural-Analytical Problems with the Help of Data Bases on Molecular Spectroscopy (MS, IR, NMR). *Chem. Sustainable Dev.* **1995**, *3*, 249–265.
- (9) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Elucidation – A Spectroscopist's Dream Comes True. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221–227.
- (10) Munk, M. E. The Role of NMR Spectra in Computer – Enhanced Structure Elucidation. *Computer – Enhanced Analytical Spectroscopy. Vol 3*; Jurs, P. C., Ed.; Plenum Press: New York, 1987; pp 127–147.
- (11) Warr, W. A. Computer Assisted Structure Elucidation. Part 2 Indirect database approaches and established systems. *Anal. Chem.* **1993**, *65*, 1087–1095.
- (12) Gloor, A.; Cadisch, M.; Kocsis, T.; Schaller R. B.; Hediger, H.-J.; Clerc, J. T.; Pretsch, E. Design criteria and implementation of hypermedia tools for structure elucidation of organic compounds with spectroscopic methods. *Anal. Chim. Acta* **1994**, *295*, 93–100.
- (13) Elyashberg, M. E.; Martirosian, E. R.; Karasev, Y. 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.
- (14) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. Application of Artificial Intelligence for Organic Chemistry: *DENDRAL Project*; McGraw & Hill, New York, 1980.
- (15) Curry, B.; Rumelhart, D. E. MSnet: A Neural Network That Classifies Mass Spectra. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213–237.
- (16) Lohninger, H.; Stancil, F. Comparing the performance of neural networks to well-established methods of multivariate data analysis: The classification of mass spectral data. *Fresenius J. Anal. Chem.* **1992**, *344*, 186–189.
- (17) Klawun, C.; Wilkins, C. L. Joint Neural Network Interpretation of Infrared and Mass Spectra. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249–257.
- (18) Eghbaldar, A.; Forrest, T. P.; Cabrol-Bass, D.; Cambon, A.; Guigonis, J. M. Identification of Structural Features from Mass Spectrometry Using a Neural Network Approach: Application to Trimethylsilyl Derivatives Used for Medical Diagnosis. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 637–643.
- (19) Varmuza, K.; Werther, W. Mass Spectral Classifiers for Supporting Systematic Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 323–333.
- (20) Haraki, K. S.; Venkataraghavan, R.; McLafferty F. W. Prediction of Substructures from Unknown Mass Spectra by the Self-Training Interpretive and Retrieval System. *Anal. Chem.* **1981**, *53*, 386–392.
- (21) Stein, S. E. Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644–655.
- (22) Neudert, R.; Bremser, W.; Wagner, H. Multidimensional Computer Evaluation of Mass Spectra. *Org. Mass Spectrom.* **1987**, *22*, 321–329.
- (23) Cone, M. M.; Venkataraghavan, R.; McLafferty, F. W. Molecular Structure Comparison Program for the Identification of Maximal Common Substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.
- (24) Lebedev, K. S.; Tormyshev, V. M.; Derendyaev, B. G.; Koptuyug, V. A. A Computer Search System for Chemical Structure Elucidation Based on Low-Resolution Mass Spectra. *Anal. Chim. Acta* **1981**, *133*, 517–525.
- (25) Scsibany, H.; Varmuza, K. Common substructures in groups of compounds exhibiting similar mass spectra. *Fresenius J. Anal. Chem.* **1992**, *344*, 220–222.
- (26) Lebedev, K. S.; Use of IR and Mass Spectroscopic Databases to Establish the Structure of Organic Compounds. *J. Anal. Chem. – Engl. Tr.* **1993**, *48/5*, Part 2 (May), 603–611.
- (27) Kalchauer, H.; Robien, W. CSEARCH: A Computer Program for Identification of Organic Compounds and Fully Automated Assignment of Carbon-13 Nuclear Magnetic Resonance Spectra. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 103–108.
- (28) Lebedev, K. S.; Scharapova, O. N.; Korobeinicheva, I. K.; Kokhov, V. A. Large Structural Fragments Determination of an Unknown Using a IR Library Search System. *Sib. Khim. Zhurn.* **1993**, *1*, 50–56 (Russ).
- (29) Kwok, K. S.; Venkataraghavan, R.; McLafferty, F. W. Computer-Aided Interpretation of Mass Spectra. III. A Self-Training Interpretive and Retrieval System. *J. Am. Chem. Soc.* **1973**, *95*, 4185–4194.
- (30) Gray, N. A. B.; Carhart, R. E.; Lavanchy, A.; Smith, D. H.; Varkony, T.; Buchanan, B. G.; White, W. C.; Creary, L. Computerized Mass Spectrum Prediction and Ranking. *Anal. Chem.* **1980**, *52*, 1095–1102.
- (31) Stokov, I. I.; Lebedev, K. S. A New Modular Architecture for Chemical Structure Elucidation Systems. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 741–745.
- (32) Molodtsov, S. G. Computer-Aided Generation of Molecular Graphs. *Comm. Math. Chem. (MATCH)* **1994**, *30*, 213–224.
- (33) Gasteiger, J.; Hanebeck, W.; Schulz, K.-P. Prediction of Mass Spectra from Structural Information. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 264–271.

CI970083B