Russian Chemical Reviews **68** (7) 525 – 547 (1999) © 1999 Russian Academy of Sciences and Turpion Ltd

UDC 543.42:681.3

Expert systems for structure elucidation of organic molecules by spectral methods

M E Elyashberg

Contents

I.	Introduction	525
II.	Methods for identification of structural fragments	526
III.	Generation of structural formulae of isomers	531
IV.	Prediction of spectra of probable structures	532
V.	Pathways of development of expert systems	537
VI.	Strategy of molecular structure elucidation	542
VII.	Conclusion	545

Abstract. The state-of-the-art of the investigations aimed at creating expert systems for establishing the structure of organic molecules from IR, ^{1}H and ^{13}C NMR spectra is analysed. Computer methods used for identification of molecular fragments, generation of their structure and spectra prediction are considered. Principles of the creation of modern expert systems and general strategy of solving structural problems are discussed. The bibliography includes 174 references.

I. Introduction

More than 30 years have elapsed since the publications [1–5] in which the problem of using mathematical methods and computers for establishing the structure of organic molecules from their spectra were first considered. Even in these articles it was shown that the structural-group analysis (SGA) can be formalised using methods of mathematical logic and performed on a computer,[1,2] while application of graph theory and combinatorial analysis makes it possible to generate automatically (perform mathematical synthesis) structural formulae from fragment sets.[3] In this time, many interdisciplinary sciences have united on the basis of mathematical algorithms as computer programs to form a new line of investigation, application of artificial intelligence (AI) to the determination of molecular structures from spectral data.

Since World War II, spectral methods have received wide acceptance in chemistry. It will not be an exaggeration if we say that IR, NMR spectroscopy and mass spectrometry revolutionised the methodology of chemical investigations and 'chemical' thinking and have become one of the main sources of structural information for chemists. Unquestionable are the advantages of spectral methods, *viz*., their versatility, fast operation, ease of constructing a computer interface and the use of well developed theories (this particularly concerns IR and NMR spectroscopy). It is these features, as well as the accumulation of a great body of information on empirical spectrum – structure correlations (SSC), that have stimulated investigations on the application of AI in chemistry.

To formalise the process of molecular structure determination, it was necessary to find mathematical tools adequate to the problem considered. Since a molecule has a discrete nature and its energy levels are discrete too, mathematical logic, graph theory and combinatorial analysis were chosen as appropriate mathematical tools. It is discrete mathematics which describes the models of the 'black box' type that has become one of the main instruments of mathematical chemistry, which has progressed intensively in the last decades.

At first, mathematical synthesis of the structures from the atoms and fragments seemed to be merely a stage in the identification process of organic compounds; however, soon it became clear that its importance goes far beyond the scope of the problem considered. In fact, innumerable sets of organic molecules designed by the structure generator can be interpreted as enumeration of logical implications from a system of axioms that form a basis of the theory of molecular structure. At the same time, knowing only the structural formula of a compound, one can predict a number of its important properties. This made it possible to pose a problem of the development of mathematical methods for predicting the structure of the compounds with desired properties and, *vice versa*, the properties of compounds from a structure given (studies of N S Zefirov's scientific school in this line of investigation have gained world-wide acceptance). Rich experience accumulated by chemists shows that most predictions on the possibility for unknown compounds to exist in nature and to be synthesised, made on the basis of the classical theory of the structure of organic substances, were borne out, the percentage of wrong predictions being relatively small.[6]

Formalisation of chemical research based on integration of many sciences has become an important result of the studies on the application of AI in molecular spectroscopy.

By and large, the development of concepts and methodological approach in this new area of study was completed by the early 1980s, which was embodied in a number of monographs [7–10] and reviews.[11–15] The efforts of investigators from many countries have resulted in general principles for creating AI systems [currently, they are called expert systems (ES)] and in developing corresponding algorithms which were tested on a large number of problems.

About twenty ES for molecular spectroscopy have been documented. They differ in their possibilities, use different algorithms and instrumental tools and are aimed at the applica-

M E Elyashberg All-Russian Research Institute of Organic Synthesis, ul. Radio 12, 107005 Moscow, Russian Federation.
Fax (7-095) 261 07 77. Tel. (7-095) 261 51 61.
E-mail: elyas@molspec.msk.ru

Received 25 February 1999
Uspekhi Khimii **68** (7) 579 – 604 (1999); translated by A M Raevsky

tion of different methods of spectral analysis. However, common to these systems is the fact that they simulate reasoning of an expert (spectroscopist) when solving an identification problem. This thinking process involves the following main stages:

(1) structural-group analysis of an unknown compound;

(2) generation of possible structural formulae from the fragment sets found;

(3) test for the correlation between structural formulae and experimental spectra (and additional information) using spectra prediction of the assumed structures.

Each stage is simulated by a corresponding block of an expert system. The above-mentioned blocks are relatively independent and their functions are usually realised using different mathematical methods. Currently, methods for identification of the fragments, structure generation and prediction of their spectra have become almost independent divisions of computer spectroscopy and mathematical chemistry.

To come as close to solving actual problems using ES as possible, it was necessary to investigate the nature of the structure determination problem. Peculiarities of this problem were considered [16] as an example of the inverse problem.[17] As is known, the inverse problems belong to the so-called ill-posed problems: they do not have a unique solution unless *a priori* restrictions are imposed.[18] Important conclusions were drawn that it is impossible to perform a structure determination without the attention of a chemist and that all problems should be solved in the interactive mode. At that time, it was also appreciated that information on the spectra – structure interrelations is usually 'fuzzy' (ambiguous). With this in mind it was suggested [19] that methods of fuzzy mathematics [20] should be used for SGA formalisation and approximate spectra prediction. The results obtained by different research groups up to the early 1990s have been reported in reviews.[21–23] At about the same time a considerable interest has been expressed by investigators in algorithms of artificial neural networks (ANN),[24–27] the application of which in different areas of computer chemistry seems to be rather promising. The number of studies in this investigation line grows rapidly. It appeared to be so new for chemistry and chemometrics that no mention of ANN was made, *e.g.*, in a 'Chemometrics' monograph [28] published in 1986.

At the same time it became clear that ES have a chance to become an analytical tool of organic analysis only if they 'learn' to operate on large non-additive structures (*e.g.*, the structures of natural compounds) and possess a high degree of user-friendliness. The efforts of several research groups focussed on this line of investigation have led to the creation of new ES (see, *e.g.*, Refs 29–32).

In the present review, the experience mostly accumulated in the last decade in the area of creation of the ES oriented to the molecular structure determination from IR, ^{1}H and ^{13}C NMR spectra is generalised. First, methods for identification of fragment sets, structure generation and prediction of their spectra are considered and then strategies of structure determination are analysed taking the most perfect ES as examples. Particular emphasis is placed on discussing the problems that arose in the course of ES operation and analysing possible ways of overcoming these problems.

II. Methods for identification of structural fragments

To identify the structural fragments, most ES use characteristic features of functional groups and other structural elements in the molecular spectra. Usually, the knowledge base of an ES consists of molecular fragments, ranges of changes in their characteristic features (frequencies in IR spectra and chemical shifts in NMR spectra) and the rules relating the spectra to the structure. In a number of systems, libraries of the structures and corresponding spectra are included in the knowledge bases in addition to molecular fragments. Depending on the type of the knowledge base, different methods of structural information retrieval are used.

1. Formal logic method

It has been shown [1,33] that the scheme of relations between the elements of molecular structure and their spectral features can be described by a discrete structural model, therefore application of mathematical logic in the studies of the spectrum – molecular structure interrelation naturally follows from the nature of the object studied. It is assumed that the molecule is additive or partially additive and each fragment A_i of a set A of fragments related to the problem given is characterised by one or several intervals of changes of characteristic features $\omega_i \in \Omega$ (Ω is a set of spectral features).

Boolean algebra was used as a mathematical tool of the theory of SGA. This made it possible to derive in the general form logical relations connecting structural elements to their characteristic spectral features and to construct a model of reasoning when identifying structural fragments.

Boolean algebra deals with elementary propositions and their combinations, which are called Boolean functions. Combinations of elementary propositions are constructed using 'and' (conjunction), 'or' (disjunction), negation and implication operations. The latter has a meaning of a complex proposition 'if..., then ...' and is denoted by an arrow ($\rightarrow$). For instance, the sentence 'if a molecule contains a fragment A_i, then feature ω is observed in its spectrum' is represented by the $A_i \rightarrow \omega$ implication. Elements of Boolean algebra and SGA theory are reported in the monograph.[7]

In the general case, reasoning of an expert who performs SGA seems to be as follows: if a Boolean function T(A, Ω) correctly reflects interrelation between the spectrum and molecular structure, then the presence in the spectrum of a combination of spectral features, which is described by a Boolean function $R(\Omega)$, indicates the presence in the molecule under study of at least one combination of the structural elements defined by a Boolean function f(A), *i.e.*

$$T(A, \Omega) \rightarrow [R(\Omega) \rightarrow f(A)].$$

This expression is a general formulation of the problem of qualitative spectral analysis in terms of Boolean algebra, it is a logical equation with respect to function f(A). Solution of the logical equation is reduced to calculating the function f(A) using an algorithm which is based on the concept of designation numbers and reported in the monograph.[7] Function f(A) contains fragment sets, each of which can explain characteristic features in experimental spectra.

The approach described serves as the basis for the SGA block in several generations of the RASTR system [34–36] and is used in the X-PERT,[29] CHEMICS,[37] EXPIRS [38,39] and other systems. This makes it possible to select successively possible fragments for producing logically consistent sets appropriate for further structure generation. However, such an approach has a drawback consisting in too 'rigid' logical determinacy. For instance, a small deviation of a characteristic feature of the fragment in the experimental spectrum from the limits of a given interval leads to the loss of the corresponding fragment. In addition, the knowledge base of the algorithm remains to a great extent similar to correlation tables and, hence, vast information presented in terms of natural language in numerous monographs on molecular spectroscopy (regularities, empirical rules, *etc.*) is not used. In this connection attempts have been undertaken to create such an ES in which SSC are described in more detail.[19,40]

2. Methods of predicate logic

To create an ES which operates on the knowledge presented in terms of a natural language with professional restrictions, it is necessary to find appropriate mathematical tools. The requirements placed on such mathematical tools are caused by the character of the knowledge formulated in terms of natural language. On the one hand, descriptive (declarative) knowledge,

in particular, in the field of IR spectroscopy (see, *e.g.*, Ref. 41), is characterised by some fuzziness and ambiguity. The statement that, usually, stretching vibrations of the C=C bond are observed in the IR spectrum in the region near 1650 cm^{-1}; however, sometimes they are shifted to the region 1685 cm^{-1} can serve as an example.[41] On the other hand, logical inferences in the course of solving qualitative problems are also performed in terms of a natural language with inherent ambiguity and hidden estimates of reliability (*e.g.*, 'often', 'rarely', 'sometimes', *etc.*). The man thinks using fuzzy concepts rather than numbers. Therefore, theoretical tools appropriate for adequate simulation of our knowledge and reasoning must meet the requirement figuratively formulated by L Zadeh who initiated fuzzy mathematics:[42] 'The theory of nature must reflect the fact that nature paints by arbitrary strokes rather than by a ball pen'.

To represent the knowledge on IR spectroscopy, applied logical calculus[19,40] (ALC) based on the theory of first-order predicates[43] and theory of fuzzy sets[20,42] was proposed. The basic principles of ALC are as follows.

A logical function $P(X_1, X_2, ..., X_n)$, where $X_i \in M(X_i)$, which takes a 'true' or 'false' value on substituting the objects belonging to a certain object domain is called an *n*-place predicate. Most often, the predicate reflects properties of the objects and relations between them. For instance, let predicate $P(X)$ reflect the property of the substances to have a strong IR absorption band near 1700 cm^{-1} and let X_1 = {acetone} and X_2 = {propylene}. Then, obviously, $P(X_1) = 1$ ('true') and $P(X_2) = 0$ ('false'). If $M(X_i)$ is a set, then a fuzzy subset F of the set $M(X_i)$ is defined as a set of ordered pairs $\{X_i, \mu_A(X_i)\}$, where $\mu_A(X_i)$ is a characteristic membership function, which varies from 0 to 1 and indicates to which degree the element X_i is a member of the subset F. For instance, the membership function of the proposition 'there is a band near 1710 cm^{-1} in the IR spectrum' will be written as follows: $\mu_A(1710\ cm^{-1}) = 1$. In this case it is assumed that there exist certain bounds, *e.g.*, 1700 and 1720 cm^{-1}, and beyond them the function $\mu_A(X_i)$ is zero.

If predicate variables are assigned the meaning of the elements of fuzzy sets, then we get fuzzy predicates. Each *n*-place predicate symbol is associated with mapping M^n into a set of the values of a characteristic membership function, which in the case of ALC is called a truth function μ_p, $0 \leqslant \mu_p \leqslant 1$.

A system of predicates describing the IR spectrum – structure interrelation was developed to describe the spectrochemical knowledge. In particular, predicates characterising the IR spectrum ('there is a band of vibrations of the type ζ of a structural group *g* in the spectrum; the band is characterised by frequency X_i, intensity Y_i and halfwidth Z_i') and molecular structure ('the molecule contains a fragment R_1', 'the molecule contains a fragment R_1 separated by n_1 bonds from m_1 fragments R_2', 'the molecule contains a ring of size *j*, which includes a fragment R_3', *etc.*) were introduced. Specific predicates characterise the state of aggregation of the substance, solvent type, *etc.* To describe changes in the data, second-order predicates are introduced, and using them it is possible to indicate the direction of changes in the variables which appeared in the first-order predicates (*e.g.*, 'increases', 'decreases', 'disappears', 'becomes split into *d* components', *etc.*).

Using such a system of predicates and conventional logical operations, it is possible to formalise declarative knowledge to a great extent. For instance:

(1) 'stretching vibrations ν_s of compounds with one double C=C bond and alkyl substituents (Alk) absorb in the region 1640 – 1680 cm^{-1}'

$$STRUCT(C=C, Alk) \rightarrow SPECTR(\nu_s, C=C, 1640, 1680);$$

(2) 'conjugation of double bonds of two alkyl-substituted groups results in splitting of the band of stretching vibrations ν_s into two bands'

$$Q_1[STRUCT, C=C, Alk(C=C)] \rightarrow \rightarrow Q_2[SPECTR, \nu_s, C=C, X_1(SPL,2)].$$

The logical formula G and the predicate symbol P are associated with a set of outside values of the truth function μ_e, $0 \leqslant \mu_e \leqslant 1$. The external value of the function μ_e reflects the knowledge quantification ('often', 'sometimes', 'in some instances', *etc.*). For instance, the statement 'stretching vibrations of the C=C bond are often observed in the region 1640 – 1660 cm^{-1}' is represented by elementary formula [0.8]SPECTR(ν_s, C=C, 1640, 1660 {10}), where [0.8] is the external value of the truth function, {10} is the segment length (in cm^{-1}) on both sides of the bounds of the region 1640 – 1660 cm^{-1}, on which the trapezoidal membership function decreases to zero. To determine the value of the truth function μ_i appeared in the elementary formula, it is necessary to multiply the external value of μ_e by those of the membership functions of fuzzy constants and variables in this formula. In this case, taking into account the properties of a triangle, we get $\mu(X) = 0.5$ at $X = 1665\ cm^{-1}$; at $\mu_e = 0.8$, the μ_i value will be equal to $0.5 \times 0.8 = 0.4$.

Spectrochemical knowledge can be divided into facts (expressions containing constants) and regularities (expressions containing at least one logical variable). It is convenient to operate on structurised knowledge. To structurise the knowledge, a class is specified of general regularities that are valid for most organic molecules (*e.g.*, 'there are no small rings containing a triple bond in organic chemistry'). Principal functional groups possessing characteristic spectral features (C = C, C≡C, C = O, C≡N, OH, *etc.*) are accepted as structurisation units of other knowledge. The molecular fragment is described by its core (C = C, C≡C, *etc.*) to which surrounding atoms are successively added, thus respectively forming first-, second-, third-level fragments, *etc.* The knowledge hierarchy thus formed is a tree graph called a declarative knowledge network. The description of a network node contains the fragment structure, possible types of its vibrations, regularities and facts for each type of vibrations and those characteristic of the given fragment.

If declarative knowledge is considered as a system of axioms of the theory used for solving an applied problem, the SGA problem is represented in terms of statements (theorems) that should be proved or rejected. It has been shown[43] that the classical resolution method adapted to the theory of fuzzy predicates can be used for automated proof of the theorems (automated SGA).

The formalism suggested makes it possible to solve SGA problems as well as those of structural interpretation of IR spectra. The results of interpretation can be used for self-learning of declarative knowledge, in the course of which the facts and regularities are tested, values of characteristic membership functions are changed and new facts and regularities are revealed.

The approach described was realised as a research prototype of an ES written in the Prolog algorithmic language.[44] The system reveals consistent fragment (network node) sets that explain the spectrum. Simultaneously, interpretation of the IR spectrum is performed: each frequency is assigned to a certain type of vibrations of corresponding fragments with an indication of the value of the truth function μ_p for the assignment given.

Predicate logic and Prolog algorithmic language were also used for structural intepretation of IR spectra in the EXPEC system.[45–47] The knowledge base of the system includes a rule interpretation generator. As in the preceding system,[44] fragments included in the knowledge base are hierarchically structurised in the form of a network of declarative knowledge. When interpreting the spectrum of unknown compound, the network is scanned to find the rule for each fragment, according to which the relation between the fragment and spectrum is analysed. Each result obtained is associated with its probability. Explanatory facilities make it possible to trace the reasoning process of the program.

The knowledge base of the system is formed on the basis of SSC known from the literature and is supplemented using the rule interpretation generator,[45] as well as due to input of expert's knowledge. To perform automated rule generation, a library of IR spectra of organic compounds belonging to the CHO class is used. The program generates a set of potential rules, selects the least correlated ones and minimises their number.

The presence of the rule generator [45] is an advantage of the EXPEC system. Automated rule generation involves four steps: (1) selection of structural fragments; (2) database retrieval to find all structures containing a S_i fragment; (3) search for the intervals of characteristic features for S_i and (4) selection of intervals appropriate for using interpretation rules. The problem of the knowledge base formation is of fundamental importance for creating an efficiently operating ES, therefore the above-mentioned approach to automatisation of this process appears to be rather promising.

The interpreter (inference machine) selects only those network fragments for which the presence of groups in the network nodes was validated. For instance, the rules concerning detection of phenol compounds can be used only after the presence of benzene ring and hydroxy group was confirmed. For each substructure, the probability of its presence is calculated using a trapezoidal probability distribution function for a given interval (*e.g.*, this shape was chosen for the band membership function of the absorption region [19, 40]). Selected fragments are arranged in order of decreasing the probability of their presence.

The investigators from different research groups suggest different rules for selecting possible fragment sets (see, *e.g.*, Ref. 48). It seems that these studies will progress and make it possible to come right up to creating ES for the molecular spectral analysis that interact with a chemist in terms of a natural language with professional restrictions. However, on this way the investigators will inevitably face many difficulties. Overcoming them will require joint efforts of the experts who carries out interdisciplinary research such as, *e.g.*, mathematical linguistics, mathematical logic, knowledge technology, *etc.*

3. Application of artificial neural networks

Beginning in the early 1990s, the attention of chemists was drawn to the possibilities of new promising mathematical tools of computer chemistry, *viz.*, artificial neural networks (ANN). Particularly rapid is the increase in the number of studies on the application of ANN to interpretation and classification of spectral data.

A neural network is a simplified model of the human brain, consisting of several layers of neurons that send signals to other neurons depending on the input signals received. Such networks function according to the 'black box' type and possess a common ability to construct empirical models of the systems for which theoretical dependences between the input and output are too complicated or even unknown. Models are obtained as a result of the network training. In the course of training, the network is represented in the form of input – output pairs related by simulated transformation. A network trained on these examples is capable of predicting the output signals from the input signals not 'seen' at that instant. The training procedure may be time-consuming (tens of hours); however, a network once trained generates the answer or prediction almost instantaneously. For instance, a network can be trained to generate structural information (output) retrieved from a spectrum (input) or to predict a spectrum (output) from structural information (input). By applying this procedure to the spectra, it is possible to establish which functional groups can be present and absent in the specimen and to complement these data with reliability estimates of the results obtained.

Let us consider general principles of designing neural networks [24, 49] in more detail.

Strictly speaking, ANN is a computational structure consisting of a number of interconnected elementary units called nodes or neurons. In a simplified form, a neuron functions as follows.

The 'signal' values s_i ($i = 1 \rightarrow m$) at the neuron input (we will call them inputs) are multiplied by weighting factors w_i and summed up to give the function Net:

$$\text{Net} = s_1w_1 + s_2w_2 + \ldots.. + s_mw_m\,.$$

Then, the Net value is transformed into the output signal of a neuron using a nonlinear transfer function. The sigmoid function

$$\text{sf} = \frac{1}{1 + \exp(-s_iw_i)}$$

is most often used as a transfer function.

This signal is then used as the input signal for other neurons and the organisation of the connections between neurons defines the network architecture. A large number of neurons and interrelations between them makes the network behaviour more complex. The network can be modified by 'adjustment' of its parameters, among which weighting factors are the main ones. Adjusting the network parameters in order to obtain a desired behaviour is called the network training. This process is carried out iteratively using a learning algorithm on a set of examples. At each iteration, the network outputs are compared with the reference values (examples) and, if they appear to be different, the program performs correction of weighting factors. The network is trained until coincidence of outputs with references.

In the first studies on structural interpretation of IR and NMR spectra, multilayer ANN trained using the back-propagation of errors (BPE) algorithm [24] were most often used. In these networks, each neuron is connected to all neurons in the preceding and succeeding layers. The first (input) layer receives the input data and distributes them over the succeeding layer. The last (output) layer, or decision layer, is responsible for outputs of the entire network. Other layers, if any, are called hidden layers (Fig. 1). Information is propagated from the input layer through hidden layer(s) to the output layer. For instance, an IR spectrum encoded as the intensity values taken at points uniformly distributed along the abscissa axis is used an the input vector. The output vector coordinates are certain numerical values (often, zeros and unities) indicating the presence or absence of particular structural features. The number of points used for encoding the spectrum defines the size (the number of nodes) of the input layer while the number of structural features (fragments) introduced determines the size of the output layer. There is no generally accepted rule for setting the number of nodes in hidden layers: it has to be large enough to provide correct data processing; however, increasing the number of nodes increases both the learning time and risk of the network 'over-training'. The latter consists in decreasing the network performance after a certain

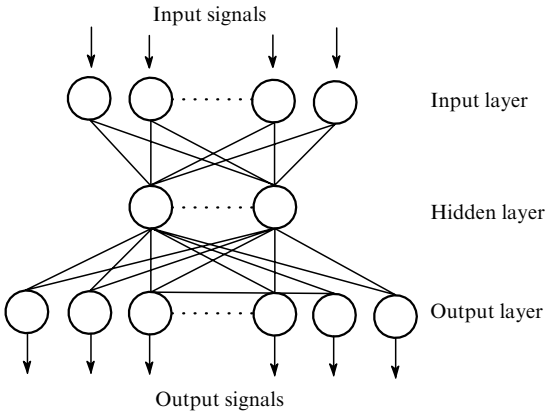


Figure 1. A scheme of neural network.

number of iterations (they are called cycles or epochs) in the course of training.

After random initialisation of the initial values of weighting factors w_i, the network training is performed on a set of spectra of the compounds containing fragments associated with desired outputs. (This set is called the training set.) There are numerous variants of this algorithm; however, all of them are based on the same principle, a steepest-descent minimisation of the total error, calculated as the sum of differences (or the squared differences) between the network outputs and desired results.

To determine the network performance (efficiency), a set of examples (the test set) is used. This set should differ from the training set in order to measure the network capacity for generalisation instead of learning.

The size and composition of the training set strongly affects the learning efficiency. For instance, increasing the number of examples increases the overall performance of the network, but only up to a certain point.[50] The effect of the composition of the training set appears to be much more complicated. The effect of presents/absents balance for structural features in the molecules, *i.e.*, the proportion of compounds presenting a structural feature to those not presenting it in the training set has been studied.[49, 51, 52] In particular, it has been found [49] that increase in the proportion of presents in the training set increases the probability of recognising these fragments in the compound analysed to the detriment of the absents. At the same time, it was found that there is a certain limit of presence under which the network is unable to perform correct classification. The setting of a well-balanced and diversified training set can be difficult, especially if the network to be trained has many outputs.

The number of network outputs is of great significance: according to the reported data,[50, 53] mono-output networks perform better than the networks with many outputs; moreover, large networks are more difficult to optimise than small ones (networks with smaller number of outputs). Since the best result for each output may occur at a different training time, the ANN training is stopped at a stage where the averaged (but not for each output) network performance will be optimum. The use of many mono-output networks for the determination of a large number of structural features is much more time-consuming than the use of a single network. Therefore, advantages and drawbacks of using a single multi-output network are opposite to those of using several mono-output networks.

A hierarchical system for spectrochemical studies, which should retain the best of both approaches, was proposed.[54] In such a system, a multi-output top-level network performs a rough classification of the spectra. Depending on the structural feature considered, the hierarchy is extended to deeper levels where further refinement is performed.

The top-level network outputs activate the first-level networks and so forth through the whole hierarchy. The design of a hierarchical system of neural networks can be considered as a compromise between one multi-output network and many mono-output networks. Its main characteristics are also intermediate. It was found that it is easier to balance the components of the training set for each network of the hierarchy than for a single large network (in some sense, as simple as for a mono-output network). In addition, optimisation of the learning time is easier for a smaller number of outputs, and the modular conception makes it possible to modify the system without the entire reconstruction.

A hierarchical system gives information on structural features on different levels: from general characteristics ('a compound contains a CO group') to more specific ones ('a compound contains an acidic group'), the number of networks to be activated in each particular case remaining small. For instance, there is no need to find out whether a compound is a primary or tertiary alcohol if the OH group has not been recognised by the top-level network. However, mention has been made [49] that the adjustment of a hierarchical network is a complex task and requires *a priori* knowledge and consideration of the dependence of sub-level networks on upper-level network performance.

Let us consider the results of studying a hierarchical system as an analytical tool for determination of the fragments using IR spectra.[49] The experiments were carried out using a set of 8620 compounds and their IR spectra taken from the Sadtler Bio-Rad library. Calculations were performed on an IBM RS6000 workstation. The system was trained to recognise 33 structural groups whose hierarchy is shown in Fig. 2. As can be seen, recognition of carbonyl group was carried out using two specially designed networks, one for wide-spread structural features (carboxyl, ketone, ester and amide) and another for rare structural features (acid chloride, anhydride and aldehyde).

The following network parameters were used.

Input vector. The IR spectrum in the region 500 – 3596 cm^{-1} was represented as a set of 259 points separated by increments of 12 cm^{-1}. The weighted mean of absorbance in the vicinity of each point was scaled to a 0 to 1 range and used as input.

Hidden nodes. Networks contained only one hidden layer, and the number of hidden nodes was set to 30 for all networks.

Transfer function. The classical sigmoid function was used.

Structure encoding: 0 and 1 were respectively used to encode the absence and presence of a structural feature.

Learning algorithm. All networks were trained using the BPE algorithm with a learning rate $\alpha = 0.05$ and a momentum $\mu = 0.90$ (Ref. 24).

Decision making. Since the network outputs were real values from the 0 – 1 range, two thresholds were necessary for decision about the presence or absence of each structural feature f in the molecule: accept level (presence threshold) AL_f and reject level (absence threshold) RL_f of a structural feature. If the output number is between RL_f and AL_f, the answer is considered ambiguous and the fragment is reported as 'not classified'. These decision thresholds were optimised individually for each network.

Training and test sets elaboration. The sets of examples required to train 10 networks constituting a hierarchical system were built by random selection from initial population with the constraints to reach a prescribed proportion of compounds containing preset structural features. For the top-level network,

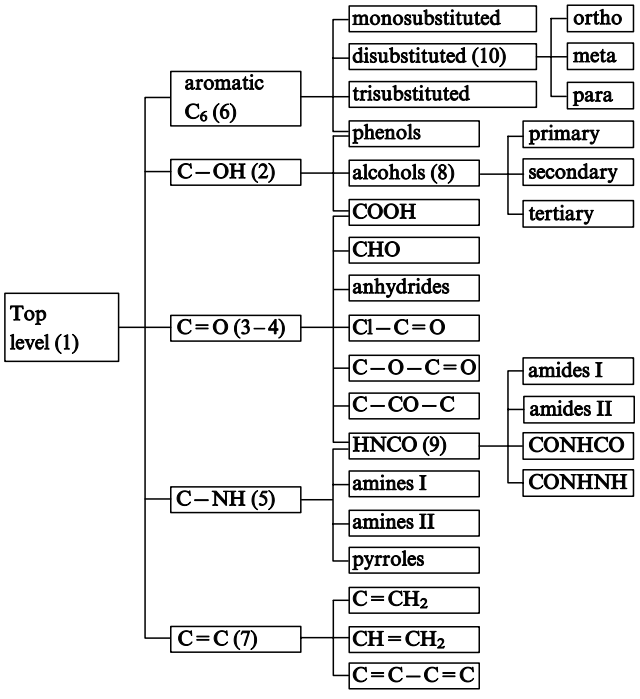


Figure 2. Hierarchical organisation of networks. The numbers of activated networks are given in parentheses.

the whole collection of spectra was used. Sub-level networks were trained using specially designed protocols.

As was mentioned above, in hierarchical structures the network dedicated to recognition of the fragments belonging to a specific class is activated only if it recognised the example as a member of this class. For a structural feature, two types of misclassification are possible: (1) the fragment was identified in the compound in which it does not occur; this case is called false present (AdP); and (2) the fragment was not identified in the compound containing it; this case is called false absent (PdA). False absents are not propagated through the hierarchy, since the structural feature is considered as absent. However, false presents raise a new problem: sub-level networks must refine classification of the fragment which actually is not contained in the compound.

Special experiments showed that the hierarchical system is robust and bears the change of networks without losing the global capability of the hierarchy. The best training was found to occur for the training sets of 1000 examples and the number of cycles was between 5 and 20. The best results were obtained for C=O and OH groups that are easily recognised in the IR spectra.

The following indices were suggested for measuring the performance of the entire hierarchical system:[49]

Ptf — a general characteristic of the presence of a structural feature (the total present found); it is defined as the ratio of the number of compounds in which the presence of the fragment was correctly identified to the total number of compounds containing the given structural feature;

Atf — the same as above, as applied to the absent structural features (the total absent found);

Pte — the total error of decision of the presence of the structural feature (the total present error), which is defined as the ratio of the number of compounds containing the feature but identified as absent to the total number of the compounds containing the given structural feature;

Ate — the total error of decision of the absence of a structural feature (the total absent error), which is defined analogously to the preceding index.

The results of testing the system using the criteria introduced are shown in Table 1. Comparison with the reported data [55] shows that, on the whole, the hierarchical system performs better than multi-output network; however, the results qualitatively rank below those obtained using well optimised mono-output networks.

Table 1. Performance of global hierarchy.

Fragment	Atf	Ptf	Ate	Pte
C–OH	93.3	88.3	6.7	16.4
CO–OH	91.2	79.8	1.0	18.0
Ph–OH	92.8	56.4	0.8	32.3
C_{sp^3}–OH	91.3	78.6	1.1	18.2
CH_2–OH	91.1	71.5	0.7	27.2
CH–OH	90.6	57.2	1.2	41.9
$C_{quatern.}$–OH	91.6	35.1	0.1	62.9
C=O	94.3	91.8	3.1	5.5
CO–OH	94.1	82.8	1.3	13.7
CO–OR	93.6	77.0	1.6	15.2
CO–R	92.9	50.7	1.7	39.0
CO–NH	91.5	85.5	3.0	11.8
$CO–NH_2$	93.4	87.0	2.0	10.5
CO–NHR	94.9	55.5	0.8	35.8
CO–NH–CO	95.6	74.2	0.1	21.4
CO–NH–NH	95.6	47.6	0.3	52.4
CHO	86.8	74.0	6.1	22.1
OC–O–CO	92.2	82.8	1.2	10.0
Cl–CO	94.6	47.0	1.7	48.0
C–NH	92.0	82.7	6.6	15.6
$C–NH_2$	91.1	52.4	3.0	43.0
C–NHR	93.4	36.3	1.8	60.3
CO–NH	91.6	66.0	1.3	29.8
Pyrrole	93.8	14.3	0.4	79.8
Aromatic C_6	79.0	86.4	16.3	5.9
C_6–OH	80.9	69.9	2.9	29.7
monosubstituted	80.8	54.0	2.1	35.0
disubstituted	70.0	53.3	10.7	37.8
ortho	73.9	34.9	0.2	62.8
meta	73.7	16.7	0.3	83.3
para	71.7	47.6	1.1	49.0
trisubstituted	76.4	34.8	6.0	59.1
C=C	93.6	42.6	6.3	56.9
$C=CH_2$	93.2	30.4	0.9	68.7
$CH=CH_2$	93.3	23.0	0.3	75.8
$Csp^2–C=C$	89.4	34.4	1.8	62.0

Currently, a few studies dedicated to application of ANN to the selection of molecular fragments has been reported; however, investigations in this area are gathering force. Thus, analyses of IR spectra using ANN were reported.[52–61] An ANN was used to identify amino acid residues in proteins from their 'fingerprints' in 2D NMR spectra.[62] Adjustment of the networks in the case of joint application of different spectral methods to elucidation of structural fragments was reported.[51,63] This investigation line seems to be of particular interest; therefore, let us consider the results reported in the last two articles in detail.

A study [51] was carried out in which IR and ^{13}C NMR spectra and molecular formulae of organic substances served as input signals for a neural network preliminarily trained to recognise molecular fragments. The ANN containing 512 inputs, 85 outputs and $\sim$ 50 neurons in the hidden layer was used. The training set comprised 1560 compounds for which IR and ^{13}C NMR spectra were known. A total of 128, 294 and 90 inputs were respectively specified for IR spectra, NMR spectra and molecular formulae. Such a distribution was chosen taking into account the estimate of the amount of information contained in each data type. IR spectra were encoded in the same way as in the study by Cleva *et al.*[49] ^{13}C NMR spectra were encoded by present/absent indicators of signals at intervals of 2 ppm. The indicators for signals of different multiplicity were stored in prescribed positions of the input vector (*e.g.*, encoding the chemical shifts of quartets in the spectral range 0–64 ppm was performed using 32 intervals to which positions 128–159 of the input vector corresponded). The molecular formula was encoded by the numbers of present types of atoms, the degree of unsaturation and molecular weight. Structural formulae were encoded by 0 or 1 depending on the presence or absence in the molecule of each of 85 fixed structural elements including functional groups ($-O-\underset{|}{C}=O$, $C\equiv N$, $H_2C=CH$, *etc.*) and chemical classes (steroids, terpenes, pyridines, *etc.*). Optimisation of parameters and the network training using the BPE procedure were carried out on HP 9000 RISC and other workstations. Performance of the system was studied for different combinations of input data.

It was established that performance of the system in the case of using NMR spectra only is higher than that reached using IR spectra only; the inclusion of molecular formula increases performance of the system for any combination of IR and NMR spectral data; simultaneous use of both types of spectra leads to a substantial increase in performance of the system (*i.e.*, the number of reliably identified groups increases appreciably).

These conclusions were not unexpected; however, the importance of this study [51] is that the procedure for identification of fragments by joint use of IR and ^{13}C NMR spectra, based on ANN application was first proposed. This approach can be realised with ease in a fragment selection block of an ES.

Application of ANN to joint interpretation of IR and mass spectra (MS) of gases has been described.[63] The network was trained to identify 26 functional groups. The proportion of correctly identified present groups was 86.4% when using com-

bined data; this index was respectively 88.4% and 78.2% when IR spectroscopy and mass spectrometry data were used separately. Among 26 groups, only for 8 did the use of MS lead to improving the recognition quality; the results of determination for the C–Cl bond have been considerably improved. This confirms the known fact that IR spectroscopy is more appropriate for identification of functional groups than mass spectrometry. Absent functional groups were predicted to an accuracy of 95.5% using combined data and IR spectroscopy data only and to an accuracy of 87% using mass spectrometry data only. The results of this study [63] show that principles of combining the data obtained by different spectral methods, as applied to ANN, require special investigations to find optimum variants.

It should be noted that, as a rule, different groups of researchers use different indices of network performance, therefore comparison of the results obtained in their studies is difficult. In this connection the problem of the selection of generally accepted indices and standard test sets, which would make it possible to easily compare the efficiency of different classification methods, appears to be topical.

Despite improved capability of ANN to identify fragments, successful application of the networks is to a great extent dependent on the user's skill in adjusting the network parameters (decision thresholds, composition of the training set, *etc*.) to meet specific requirements when solving any problem. Currently, there is no information on the use of ANN in any actual ES dedicated to identification of the molecular structure and the number of structural features that can be recognised using an ANN is unclear (for comparison, it should be noted that spectral libraries of known ES contain hundreds of fragments). Therefore an estimate of actual possibilities of this approach is still ahead. However, it can be stated with certainty that application of ANN to the determination of individual classes of chemical compounds is rather efficient.

4. Methods for recognition of large fragments

To solve the problem of identification of large molecules, an ES must have adequate analytical tools for determination of large fragments. Attempts to create libraries containing large fragments and ranges of variations of their characteristic spectral features are of no promise because of imminent combinatorial explosion and the virtual impossibility to predict the ranges. An actual way of solving the problem is using modern databases containing structural formulae and spectra of substances. The basis of the method is a hypothesis that similar spectra (subspectra) correspond to similar structures (substructures).

Two approaches to identification of large fragments using IR, NMR and mass spectra databases have been described. The algorithm suggested to this end in the first approach is reduced to three main operations:[64–72] (1) search for structures whose spectra are similar to a given experimental spectrum and ordering the former according to decrease in similarity; (2) selection of a number of structures in the beginning of the list that, according to the basic hypothesis, have the highest resemblance to the structure of the analysed compound; and (3) selection (among these structures) of largest common fragments that can be plausible constituents of the structure of the unknown compound. The probability of correct identification of the fragment increases if the database retrieval system selects the data using several types of spectra and the selected fragments are tested on the basis of coincidence of the predicted subspectra (the latter can be most easily done using ^{13}C NMR spectra).

Realisation of the above algorithm requires reliable quantitative criteria for structural and spectral similarity, as well as methods for revealing the largest common subgraph in a set of chemical graphs. Studies on the development of such methods gave rise to an independent investigation line in graph theory and numerous studies on the subject have been reported.

The other method for recognising large fragments found a wide application in practice due to the appearance of large ^{13}C NMR spectra databases. In these types of databases the structures are described by 'spherical' code [73] while the signals in the spectra are assigned to corresponding carbon atoms. This made it possible (see, *e.g*., Refs 30, 31, 74) to develop algorithms which can automatically create the subspectrum–substructure correlation libraries used for solving the problem raised. First, the program compares all subspectra of library fragments with the experimental NMR spectrum of the analysed substance. In this case it is necessary that the signal intensity (*i.e*., the number of carbon atoms) be also indicated in addition to the chemical shifts and signal multiplicities. As a result, only those substructures (fragments) are selected to which subspectra coinciding (within the limits of prescribed deviations) with the query subspectra correspond. Further, the most probable fragments are selected by a detailed analysis of spectral information and ranked in descending order with respect to the number of carbon atoms and skeletal atoms. Thus, the largest fragments appear in the beginning of the list (the use of the described procedure in ES is considered below).

III. Generation of structural formulae of isomers

No more than two dozen of publications concerning the search for ways of increasing the software performance and methods of more efficient use of structural constraints have appeared after the first generators of isomers [3, 75–78] were developed. This relatively small number indicates stubbornness of the problem rather than a loss of its actuality or interest. The numbers of possible structural formulae and the CPU time grow exponentially as the number of skeletal atoms and fragments involved in the generation process [they are called discrete units of structure (DUS)] [7] increase. Therefore, all the efforts were aimed at using structural constraints even in the course of rather than after structure generation and at developing new algorithms that make possible improving the efficiency of generation.

This interesting and ambitious problem drew the attention of mathematicians, specialists in the field of graph theory and combinatorial analysis, which resulted in the development of new approaches and appearance of efficient software (COCOA,[79] MOLGEN,[80–82] and GENM [83–86] programs). Parallel to the search for efficient generation algorithms, computer performance has increased swiftly. On the whole, this has led to appreciable increase in dimensionality of the problems which can be solved using modern generators of isomers. Currently, the number of DUS used in the structure generation that takes a 'reasonable' time lies in the range from 10 to 16 depending on the valence of the atoms considered as DUS, the number of free bonds in the fragments, the unsaturation degree of the molecule and several other factors (*cf*. with 7–8 DUS in the 1970s). For instance, it took 3420 s to generate 4347 isomers of composition $C_{15}H_{32}$ using (the first in the USSR) program generator written by V V Raznikov in codes of an M-20 computer,[75] whereas a modern GENM program installed on a personal computer equipped with a Pentium CPU solves this problem in a matter of seconds. One more example can be given: generation of 217 isomers of composition C_6H_6 using the MAISS program written in Fortran algorithmic language and installed on a Minsk-32 computer took ~30 min;[77] the GENM program performs this task for a fraction of a second, which means that the actual rate of structure generation has increased by several orders of magnitude. However, since recognition of the structures of large molecules is of prime interest for practice, the generation stage remains a bottleneck for all ES despite the advances achieved.

Basic requirements for the structure generators have been formulated and realised as long ago as 1970s. The results of these studies were generalised.[7, 8] The most important requirement is that, irrespective of generation algorithm used, the software suitable for practical use must contain flexible tools for imposing structural constraints among which the following are typical.

1. Fragments — DUS identified at the SGA stage. It is usually assumed that they have no overlapping atoms and the number of

free valence bonds of terminal atoms in such fragments is fixed. Therefore such fragments are also called macroatoms and they participate in structure generation along with free atoms. The DUS sets are designed of atoms and macroatoms and the total number of DUS in the set determines the dimensionality of the problem. Some programs, *e.g.*, GENOA [87] admit partial overlap of the macroatoms; however, this makes the algorithm more complicated and increases the generation time.

2. Obligatory and forbidden fragments (GOODLIST and BADLIST, respectively). These are structural constraints imposed by a chemist and each structure must meet these requirements. This type of constraint is imposed on the basis of *a priori* knowledge of the compound under study (the origin of the specimen, data of non-spectral experiments, *etc.*). These fragments have a peculiarity consisting in the possibility for their structures to overlap and superimpose on the DUS structure.

3. Constraints associated with the ring size and bond multiplicities make it possible to formulate general requirements for generated structures with consideration of specificity of the DUS set and *a priori* information (the presence of multiple bonds in macroatoms, rings, the inclusion of the unsaturation degree of the molecule, *etc.*).

4. Fragments which are unlikely in organic chemistry. These include small rings containing triple bonds, structures that contradict the Bredt rule, certain strained fused structures with small rings, fragments of the $-O-O-O-$ type, *etc.*

Since 'construction of generators' has become an individual division of mathematical chemistry and combinatorial mathematics, we will restrict ourselves to mentioning some new programs that are of prime interest for the ES creation. Detailed information concerning modern structure generation algorithms can be found in the theses.[88, 89]

The CONGEN generator developed in 1970s [90] is based on the algorithm generation by building up a chemical graph from its centre. As a result of improving this algorithm, the MOLGRAPH [80] and MOLGEN [81–83] programs have been developed. The MOLGEN program generates structures using nonoverlapping fragments. It has a high performance and is capable of generating stereoisomers. Currently, this generator has received wide acceptance in Europe.

A GENM program [84–86, 88] for generation of structural isomers was developed in the Russian Federation. The parameters of the program are at least as good as and in some instances superior to those of the MOLGEN program. The algorithm used has a number of unique features that make it possible to rule out 'hopeless' branches of the solution in the course of generation in accordance with the constraints imposed. Operating experience of the GENM generator at the Institute of Organic Chemistry (the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russian Federation) and in the RASTR [36] and X-PERT [29] systems has shown its high reliability and fast operation.

The SMOG program [89, 91] allows a chemist to analyse different structural hypotheses in the course of investigations of chemical reactions. This program can generate structures containing atoms in unusual oxidation states or carrying formal charges, generate the molecules using both non-overlapping and overlapping fragments and recognise different resonance forms of the same aromatic molecule among the results of generation. The SMOG program can serve as a basis for the development of software tools designed to operate on a special-purpose workstation used in the chemical laboratory; however, the program operates not so fast to be used as a 'conveyer-type' structure generator in ES.

In some instances where the fragments are elucidated independently by using different methods of investigation (IR, NMR, MS, *etc.*) it may appear that the structures of particular fragments have common atomic groups (*e.g.*, $-CH=CH-\underset{|}{C}=O$ and $\diagdown\!\!\diagup C=\underset{|}{C}-CH_3$). In this connection the GENOA program [87] capable of generating the structures using overlapping fragments was developed to extend the potentialities of the CONGEN program. At the same time, structure generation using overlapping atoms belonging to different fragments is the major condition for operation of a number of recently developed programs.[30, 31, 74] This type of algorithm was developed to provide structure generation using a set of fragments identified by means of ^{13}C NMR spectra (see Section II.4). The program tries to match the atoms belonging to different fragments and predict the subspectrum of a new, enlarged fragment. If the predicted subspectrum does not contradict the experimental spectrum within the limits of prescribed admissible deviation, an attempt is made to add one more fragment, *etc.*

Thus, structure generation is a successive assembly accompanied by verification of the correctness of all possible combinations of the fragments. Structure generation using this approach does not require knowledge of an empirical formula and even the molecular weight. However, it should be compensated for the absence of empirical formula by ascribing a certain number of carbon atoms to each signal in the experimental NMR spectrum (this cannot be done in all cases, especially for symmetric molecules). This algorithm implicitly uses the $C_nH_{(m-k)}$ part of the empirical formula, where k is the number of hydrogen atoms bonded to hetero-atoms. Information on the latter is retrieved in the course of structure generation. However, if the set of selected fragments is not exhaustive with respect to the structure of the entire molecule, the generation fails. Application of this type of generator is considered in detail below.

Mention may be made of a number of studies (*e.g.*, Refs 92–97) in which different principles of structure generation were evaluated. However, their actual efficiency and performance can be assessed only by solving the test problems on clone-type computers, which seems to be difficult.

In some instances, structure recognition is not restricted to determination of the structural formula of the compound, but also involves the establishment of the spatial model of the molecule. To this end, algorithms and programs for generation of stereoisomers,[83, 98, 99] and building of spatial models of the molecules [99–103] were developed. They use structural formulae of organic compounds as the input information. Building spatial structures can appear to be necessary for prediction of the spectra of candidate structures, thus becoming an important stage of elucidation of the true structural formula of the compound analysed. It should be noted that one of the most efficient 3D-generators, CORINA,[101] can be accessed *via* the Internet (the webpage URL is http://www2.ccc.uni-erlangen.de/servies/3d.html).

IV. Prediction of spectra of probable structures

Prediction of the spectra of isomers to reveal plausible structures is the final stage of the ES operation. In practice, depending on how hard are structural constraints imposed, the answer file may contain tens, hundreds or even thousands of structural formulae that cannot be distinguished taking into account the ranges of changes in characteristic spectral features. This requires prediction of complete molecular spectra. Time-consuming quantum-chemical calculations are inapplicable in this case, so only fast approximate spectra prediction using various empirical correlations, rules and parameters, as well as spectral databases can be used. Numerous studies concerning this problem have been reported. Because of its prime importance for the methodology of the development of ES, let us consider the approaches to prediction of IR, ^1H and ^{13}C NMR spectra.

1. Simulation of ^{13}C NMR spectra

^{13}C NMR spectroscopy is one of the most powerful methods for elucidation of the structure of organic molecules and testing structural hypotheses. Among all types of molecular spectra, it is the ^{13}C NMR spectra that can be simulated most easily since in the case of off-resonance they consist of singlets, doublets, triplets and

quartets ($^1H-^{13}C$ spin–spin coupling through one bond) and each anisochronal (magnetically equivalent) carbon atom is ascribed a signal with specific chemical shift and multiplicity. Usually, 2D DEPT NMR experiments make possible reliable determination of signal multiplicities even in the case of strong overlap. Prediction of ^{13}C NMR spectra can be performed with different degree of detailisation.

a. Prediction of the number of signals with different multiplicity

Shelley and Munk [104] were the first who suggested use of prediction of the number of signals with different multiplicity for selecting candidate structures. The method is based on the topological analysis of a multigraph on which the structural formula of the molecule is mapped. All skeletal atoms are divided into groups of topologically equivalent atoms. Where all carbon atoms in the structure appear to be unique, the predicted number of signals in the spectrum is equal to the number of carbon atoms in the molecule; in this case the number of different types of multiplets is predicted automatically. If the structure contains several groups of equivalent carbon atoms, the expected number of signals in the spectrum is equal to the number of groups. The authors [104] analysed how stereochemical factors affect the correctness of prediction and suggested a number of rules which make it possible to take into account splitting of the signals of dimethyl groups located near a chiral centre. A simple comparison of predicted and experimental numbers of different types of multiplets allows one to reduce appreciably the number of candidates. However, it should be noted that accidental coincidence of inequivalent signals can occur or unexpected splittings can appear in the spectra. To improve this approach, it was suggested [29] that the number of signals with different multiplicity in characteristic spectral regions be predicted, which increased the resolving ability of the method used.

b. Prediction of chemical shifts

Simulation of ^{13}C NMR spectra is based on the prediction of chemical shifts of carbon atoms in the structure under consideration. The model spectrum obtained is compared with the experimental one and the structural hypothesis is accepted or rejected on the basis of calculated mean or standard deviations. There are two most widely used procedures for predicting NMR spectra, based on construction of empirical models [105–110] and on application of spectral databases.[111–115] Recently developed software [116–118] uses both approaches. Attempts to simulate ^{13}C NMR spectra using ANN are also documented.[119–128]

Grant and Paul [129] suggested the first additive linear model for calculating the chemical shifts of carbon atoms in aliphatic hydrocarbons using increments with consideration of the effects of environment up to δ-atoms. Later, linear models were constructed for alkanes, cyclohexanes, alcohols, amines, chlorine-containing compounds, *etc*. The formulae derived and tables of increments were collected in a number of monographs (see, *e.g*., Refs 130, 131). The approach described provides a rather high quality of prediction, which stimulated the development of software for analysing the candidate structural formula, selecting a corresponding additive model for each carbon atom and calculating the chemical shift. This type of software is of limited utility since chemical shifts can be predicted only for sp^3-hybridised carbon atoms and the increment values cannot be modified.

Fürst and Pretsch [110] removed these restrictions. Using large databases containing structural information and assigned ^{13}C NMR spectra, the authors of this study succeeded in constructing linear models which can be applied to many classes of chemical compounds. The models contain configuration- and conformation-dependent parameters and take into consideration the configuration of C=C bonds and the presence of axial and equatorial substituents in the cyclohexane ring. The software developed allows modification of parameters (reference values and increments) and input of new additive rules. A test of the program on the set of 170 000 known chemical shifts showed that the rules derived are fulfilled for 97% of the cases. The standard deviation of predicted chemical shifts from experimental values was 5.5 ppm; our experience has shown that calculations of the spectrum of medium size molecules ($C_{12}-C_{16}$) on a modern personal computer take no more than a second. Obviously, this program can be used for filtration of the answer files created by an ES.

In another approach,[111–115] databases containing structural information and chemical shifts of carbon atoms are used. For all carbon atoms of each reference compound contained in the database, atom-centred fragments (ACF) with a prescribed number of concentric layers are generated. These fragments and the corresponding chemical shifts are stored as an ordered list. To predict the spectrum of a candidate structure, the program selects all possible ACF existing in the structure, performs a search for their analogues in the database and ascribes the chemical shifts taken from the reference fragments to carbon atoms. If an ACF was not found in the database, the program usually performs 'interpolation' using the available 'most similar' structures. The results obtained using this approach are in very good agreement with experimental ones when using a large database containing a diversified set of structures (see, *e.g*., Ref. 73). Usually, the difference between predicted and experimental chemical shifts lies within the limits of 1 ppm, which means that the prediction accuracy is rather high.

However, despite its indubitable advantages, the approach described has some drawbacks. The absence of stereochemical information in the database (only structural formulae are available) may strongly affect the prediction quality. As has been shown,[132] neglect of stereochemical effects may increase the difference between predicted and experimental chemical shifts by more than 10 ppm. Spectral properties of the reference fragments used for extrapolation may appear to be not the closest to those of the fragments of the hypothetical structure. Additionally, the fact that the effect of an atom equally distant from another atom can be different in different compounds is ignored at the encoding stage. Finally, prediction of a spectrum takes a rather long time (tens of seconds and even minutes) depending on complexity of the molecule.

A method is described,[117] which is a combination of both above-mentioned approaches. At the first stage, the environments of carbon atom in the reference and target (verified) structures are represented as vectors and the Euclidean distance between these vectors is calculated. After repeated application of this procedure, a test set of carbon atoms whose chemical environment is similar to that of carbon atoms in the target molecule is generated. If the degree of similarity of the chemical environment of the target atom appears to be high, the chemical shift characteristic of its 'twin' in the reference structure is ascribed to this atom in the target structure. Otherwise a linear model is constructed using the test set. To this end, descriptors describing topological, geometric and electronic properties of the environment are introduced. Models are constructed using multiple linear regression (MLR), partial least squares (PLS) method [28, 133] and principal component analysis.[28] The test of the combined approach on the set of 38 500 atoms has shown that the mean deviation is 1.69 ppm, whereas this value was 1.85 ppm in the case of straightforward database retrieval procedure.

It was mentioned above that ANN can be used not only for structural interpretation of the spectra, but also for prediction of spectral pattern corresponding to a particular structure. Attempts to apply ANN for prediction of chemical shifts in NMR spectra have been undertaken taking methyl-substituted cyclohexanes,[121] mono-substituted benzenes,[122] monosaccharides,[125] ribonucleosides,[126] sp^2- and sp^3-hybridised α-atoms in acyclic alkenes [127, 128] *etc* as examples.

Artificial neural networks are trained to predict spectra of compounds belonging to prescribed classes on a training set which includes structures encoded analogously to the procedure

described in Section II.3. In the course of training, the network uses encoded reference structures as input information and the output signals are compared with the ^{13}C NMR spectra of these structures. Training is completed if the deviations of predicted spectra from the reference ones take appropriate values. Let us consider the results of computational experiments [128] as an example.

The authors trained a network to predict the chemical shifts of sp^3-hybridised carbon atoms in the α-position with respect to the double bond in acyclic alkenes. Parallel with the training, a MLR-based model was developed. Testing both models showed that the standard deviation lay within the limits from 1.3 to 1.5 ppm for ANN and was 3 ppm for the MLR model, which means that ANN can predict the spectra more correctly than linear models.

Artificial neural networks can be applied only to predict the spectra of compounds belonging to those classes on which the network was trained, whereas the structures included in the answer (output) files created by the ES may contain diversified and unexpected atomic fragments. We believe that the optimum strategy of applying prediction of ^{13}C NMR spectra must include two steps, *viz*., screening of structures (using fast approximate prediction methods, *e.g*., linear models [110]) followed by more accurate (oriented to chemical classes) prediction applied to those structures for which the mean or standard deviation lies within the limits of error of the computational method used and, hence, further refinement is required.

2. Prediction of ^1H NMR spectra

As was mentioned above, relative simplicity of prediction of ^{13}C NMR spectra is associated with the fact that they consist of clearly defined multiplets. It is a much more complicated problem to predict ^1H NMR spectra, because they contain both first-order signals and higher-order multiplets. This is also the reason for difficulties encountered when trying to compare automatically experimental and predicted spectra. That is why information on the methods for prediction of ^1H NMR spectra is scarce.

The Proton Shift program developed [134] to predict the chemical shifts in ^1H NMR spectra on the basis of simple linear models [131] automatically determines the substructures for which certain rules are applicable and considers the rest part of the molecule as a set of substituents at each of the substructures selected. To extend the area of application of the program, different extrapolation procedures were used, including division of large substructures and replacement of missing substituents by related substructures. The program takes into account *cis*–*trans*-isomerism of double bonds as well as axial and equatorial positions of substituents in six-membered rings. Linear models are based on a large set of increments and the total number of parameters that can be used in calculations approaches 3000. A test on the set of 583 chemical shifts showed that the mean deviation of predicted values from the measured ones was 0.06 ppm (the standard deviation was 0.18 ppm).

The authors of another study [135] on evaluation of the quality of the Proton Shift system using a large test set found that the standard deviation of chemical shifts was 0.3 ppm, which is closer to actual values. Using the ASSEMBLE structure generator,[136] 24 308 isomers with 111 605 different proton environments were generated. The chemical shifts were assessed in 89% of the cases, which can be considered satisfactory. The mean time taken to predict the chemical shifts for one structure on a Macintosh Quadra 650 computer was 0.30 s. When comparing the predicted and experimental spectra, the input of the latter is performed manually as a set of chemical shifts of CH, CH_2 and CH_3 groups. The program compares each experimental chemical shift with the closest value of corresponding chemical shift in the predicted spectrum. Then the standard deviations are calculated to rank the isomers generated. The procedure for comparing the chemical shifts in high-order spectra used in the study cited [135] remains unclear. Most likely, the approach described can be used only for preliminary rough selection of structures in the answer file ranked after predicting the ^{13}C NMR spectra.

The ACD/HNMR Predictor program [137] is designed to calculate automatically the absorbance curve in the ^1H NMR spectrum from the structural formula of the molecule. The program has an internal database containing 180 000 experimental chemical shifts and 25 000 experimental spin–spin coupling constants (SSCC). The algorithm of quantification of spin–spin interactions in organic compounds used by the system is based on parameters found for more than 3000 structural fragments. Each δ_H value in the database is assigned to a particular structural fragment and each J_{H-H} constant is assigned to a pair of interacting structural fragments taking into account the skeleton bonding them. The internal database of assigned δ_H and J_{H-H} values is based on the results of analysis of 48 000 experimental ^1H NMR spectra. The δ_H and J_{H-H} values for fragments not found in the database are calculated using special-purpose algorithms.

Calculations of the spectrum are performed for the basic frequency of the spectrometer, which is considered as a parameter. Calculated chemical shifts and constants are provided with 95% confidence intervals to indicate the reliability of the calculated values. The spectrum is calculated with consideration of second-order interactions and long-range SSCC, which allows simulation of the spectra of strongly coupled spin systems containing up to eight magnetically inequivalent nuclei (if there are magnetically equivalent nuclei, then a larger number of interacting spins can be taken into account). The program recognises *cis*–*trans*-isomers of alkenes and *syn*–*anti*-isomers of amides, oximes, hydrazones and nitrosamines.

After input of the query structure whose spectrum should be predicted, the program automatically 'splits' the molecule into a set of unique fragments that are compared with the fragments contained in the internal database. If any structural fragment coincides with a fragment from the database, its experimental δ_H values are included into the set of chemical shifts of the structure. If some fragments are not found in the database, the chemical shifts are extrapolated using the parameters of similar fragments.

After composing the final lists of chemical shifts and SSCC, the program composes and diagonalises the spin Hamiltonian matrix to calculate the number of lines, their positions and intensities and to assign the δ_H values to the hydrogen atoms of candidate structure. The program generates the spectrum plot (including digitising of positions of the lines and the integrated curve). Calculations of chemical shifts, spectral contour, integrated curve and digitising of maxima take from several seconds to tens of seconds CPU time on a computer with a Pentium type CPU.

The ACD/HNMR Predictor program was developed to interpret experimental spectra and test hypotheses. However, where preliminary ranking of the structures in the answer file created by the ES was already done on the basis of ^{13}C NMR spectra prediction, only 5–10 structures placed at the beginning of the final list require additional testing. Hence, comparison of the calculated spectra with the experimental ones to select the most probable structures can be done with ease by a chemist.

The author of this review has assured himself that the approach used in this program is highly efficient when analysing the answer (output) files for many problems solved using ES.[31] In all the cases considered, it has been possible either to elucidate the true structure or to reduce drastically the number of alternative structures.

3. Prediction of IR spectra

The nature of IR spectra is such that, in principle, they can be used for selecting candidate structures. This is due to the fact that IR spectrum depends not only on which groups constitute a molecule, but also on their arrangement; in other words, the IR spectrum reflects properties of the entire molecule. Therefore, considerable

efforts of many investigators were aimed at developing methods for prediction of IR spectra of organic molecules.

Semiempirical calculations of an IR spectrum using a computer database of molecular parameters (the geometry, force constants, electrooptical parameters) can be considered as an efficient method for predicting the spectra. A corresponding database as well as computational techniques and software were developed under the supervision of L A Gribov.[138] The IR spectrum is calculated in the harmonic valence-force field approximation using a semiempirical computational procedure for calculating the band intensities.[139] The force constants and electrooptical parameters to be stored in the database are found from quantum-chemical calculations followed by solving mechanical and electrooptical inverse spectral problems. Calculated spectra are represented as spectral distributions of the absorption coefficient and can be immediately compared with the experimental curves. Calculations of the IR spectrum of a medium-size molecule take several minutes of CPU time; however, preparation of input data may take from 15 min to several hours depending on the problem complexity. Our experience shows that this approach can be successfully used not only to select preferable isomeric structures, but also to elucidate plausible stereoisomers and conformers.[140]

However, the advantages of this method for prediction of IR spectra have their counterpart consisting of at least three drawbacks: the method is not so fast as those used for prediction of NMR spectra and, hence, it cannot be used for 'conveyer-type' calculations of dozens of structures (to say nothing of hundreds and thousands of them); prediction of the spectrum cannot be performed automatically without attention of a chemist; and the method can be applied to recognise small- and medium-size molecules only.

The first two drawbacks are obvious and their influence will gradually decrease with improving the algorithms and increasing the computer performance. The third one is of crucial importance, since it is due to the nature of molecular vibrations rather than to limitations of the software performance imposed by the size of molecules (*e.g.*, a program [138] makes it possible to operate with the structures containing over 100 atoms). In the general case, the larger the molecule, the more rich its IR spectrum. This statement has a consequence consisting in a nearly uniform distribution of absorption bands of rather large molecules over the 'fingerprint' region, which makes it senseless to compare the calculated spectra of candidate structures with experimental ones. Therefore, it is appropriate to use straightforward calculations of IR spectra in those cases where it is impossible to make unambiguous decision using fast prediction methods.

Prediction of IR spectra of three isomers with closely related structures, for which standard deviations of predicted ^{13}C NMR spectra appeared to be within the limits of admissible errors of computational procedure, is illustrated in Fig. 3. The structures are ordered in accordance with decreasing similarity between the predicted and experimental NMR spectra. Comparison of calculated and experimental IR spectra shows that the calculated spectrum of the structure shown in Fig. 3 *b* is closer to the experimental spectrum than the calculated spectra of the other structures (see the region from 10 to 14 μm). The same result was obtained using the ACD/HNMR software package [137] developed for prediction of ^{1}H NMR spectra. Thus, three methods for prediction of spectra indicate the priority of the first structure (2,4-dimethylanisole).

Impossibility of using semiempirical methods for fast automated prediction of IR spectra has stimulated the search for another approaches. Dubois *et al.*[142] suggested the simplest procedure for simulating the IR absorbance curve. A database containing 700 substructures with fixed expected absorbance regions in the IR spectra was created within the framework of the DARC system. The expected intensities and band shapes are indicated for each interval. The prediction method includes the search for reference fragments in the candidate structure and

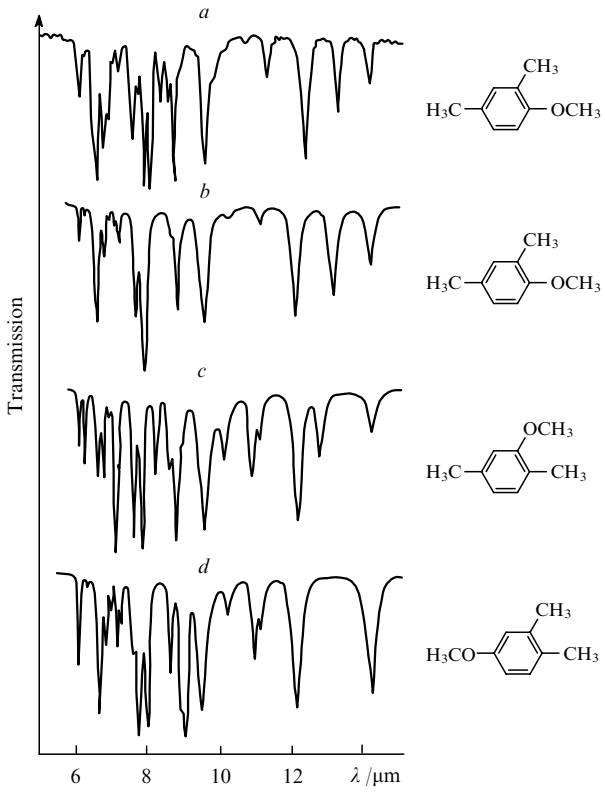


Figure 3. Experimental (*a*) and calculated (*b* – *d*) IR spectra of candidate structures.[141]

graphical representation of the spectrum of the molecule as superposition of the bands on the intervals corresponding to the fragments found. It is obvious that this approach leads to inevitable loss of many bands of skeletal vibrations in the 'fingerprint' region, as well as overtones and combination bands, and that the influence of molecular symmetry on the band intensities in the region of functional groups is ignored. It is also clear that the fragment library is far from exhausting all possible atomic combinations.

It should be kept in mind that, as a rule, characteristic spectral features of isomers included in the answer (output) file of ES correspond to experimental spectra. However, structurally similar isomers can be recognised first of all using noncharacteristic IR bands containing encoded structural information and observed in the 'fingerprint' region'. In practice, the most likely structure is chosen on the basis of careful comparison of frequencies and intensities of no more than five bands in the experimental and predicted IR spectra (see Fig. 3). Therefore the approach suggested [142] should be considered unacceptable for the analysis of the answer structural files.

An attempt to use ANN for predicting the IR absorbance curves of organic compounds from numerically encoded structural formulae has been undertaken.[143] It has been possible to predict satisfactorily absorbance in the regions characteristic of functional groups; however, this approach appeared to be unacceptable for prediction of the 'fingerprint' region. It is likely that the failure can be explained by application of too rough a coding procedure, which, in particular, completely ignores the spatial structure of the molecules.

An analogous conclusion was also drawn by Weigel and Herges [144] who attempted to simulate IR spectra using an ANN trained on a set of spectra obtained from experiments and semiempirical quantum-chemical calculations.

In contrast to the above-mentioned failures, a new approach [145, 146] seems to be encouraging. Taking into account

that IR spectra are strongly dependent on the spatial structure of molecules, the authors attempted to train the ANN on a training set composed of three-dimensional models of organic molecules. Since all models should be described by vectors of equal dimensionality, the impossibility of using Cartesian atomic coordinates in this case is obvious. To describe the spatial structure of molecules with different numbers of atoms using the same number of descriptors, it was subjected to a transformation based on the equation used for analysing experimental data on electron diffraction. This made it possible to find a procedure for coding any spatial structure by a vector with 32 or 64 coordinates (the 3D MoRSE code). Using the CORINA program [101] for fast transformation of structural formulae into three-dimensional models, the authors studied the spatial structure – IR spectra correlations using ANN. The structural formulae from the training set and the target structures were first transformed into spatial models and then encoded using the 3D MoRSE code.

Experiments show that this methodology makes it possible to simulate adequately the IR spectra of specified classes of compounds throughout the absorption region including the region of 'fingerprints'. The experimental and predicted IR spectra of a quinoline derivative are shown in Fig. 4 as an example (in this case, the ANN used for prediction was trained by the BPE method on a set of more than 50 quinoline and isoquinoline derivatives). The similarity of experimental and calculated spectral curves for this non-additive complex molecule shows that the approach has a chance to become an analytical tool for choosing plausible structures. However, it should be kept in mind that in this case successful results were obtained after training the network on the set of related compounds.

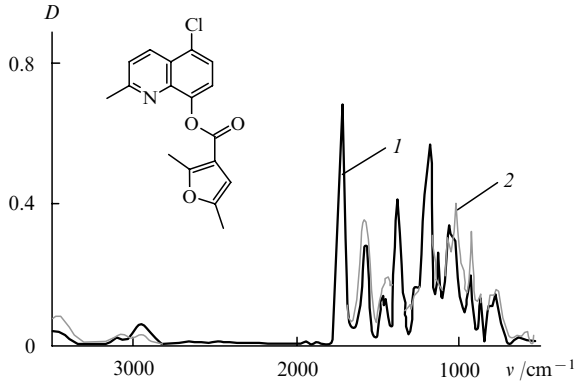


Figure 4. IR spectra of trisubstituted quinoline: experimental (*1*) and predicted using ANN (*2*).[146]

It is noteworthy that, in principle, this approach also makes possible solving the 'inverse' problem, *viz*., construction of the spatial model of a molecule using its IR spectrum. This was illustrated taking 1-phenylbutan-1-ol [146] and cholesterol [147] molecules as examples (in the latter case, the model found from the IR spectrum appeared to be fairly close to the structure obtained from semiempirical quantum-chemical calculations).

Yet another method [148] for prediction of IR spectra seems to be promising. On the basis of the search for nearest neighbours of a target molecule in the library containing structures and corresponding IR spectra the authors proposed to develop a method for simulation of IR spectra in the region 2250 – 550 cm^{-1}. The prediction mechanism is based on a reasonable assumption that similar IR spectra correspond to similar structural features. Thus, it is assumed that it is highly probable for compounds assigned the highest similarity rank to have IR spectra similar to that of the target molecule. The predicted spectrum is composed of the spectra of the neighbours (the nearest neighbours in the structure space) if they meet specific requirements for structural similarity. The structures and spectra are considered as such, *i.e*., no substructure – subspectrum correlations are used. This approach is most likely to retain noncharacteristic skeletal vibrations, overtones and combination bands, as well as the bands that are changed depending on the molecular symmetry (usually, these bands are in correlation tables). This method of nearest neighbours has also the advantages that it requires no preliminary classification of reference substances and allows easy elucidation of the target molecules that have no analogues in the reference set, thus indicating impossibility to simulate the spectrum.

Implementation of this idea required choice of such procedures for coding fragments, structures and spectra that would make it possible to introduce quantitative measures of structural and spectral similarity. Two methods of structure representation used in the study cited [148] were the method of ACF with one atomic layer [149] and a method which employs a modified graph-theoretical descriptor designated as the ES vector.[150]

The ACF code is a reasonable compromise between the undue increase in the number of fragments and comprehensive description of the molecule. In this case, each atom is characterised by its own type (C, O, *etc*.) and by the types of the atoms and bonds of its nearest neighbours (hydrogen atoms are ignored). The fragment size varies from two atoms and one bond (a terminal group) to five atoms and four bonds for the atoms with four valence bonds. The fragments are ordered in the same way as the coordinates of the *n*-dimensional vector and each molecule is characterised by its own vector. The presence or absence of a fragment in the molecule is represented as a binary 1 or 0, respectively, and the occurrence number of the given fragment in the structure is indicated.

The program selected 180 different fragments in the set of about 9000 compounds for which the IR spectra were known. The structures were coded as vectors with $n = 180$. The molecules consisting of C, N, O, Br and Cl atoms and containing ordinary, double, triple and other bonds were selected. On the average, the structure was described by eight different fragments while the reference molecules were polyfunctional.

Using the ACF code, it is possible to exactly describe various functional groups and, at the same time, impossible to distinguish the variants of mutual orientation of functional groups within the same molecule (*e.g*., substitution types in benzene ring). Therefore, the ES vector [150] was used as an additional descriptor to simulate the structure in more detail. It is less sensitive to changes in the atomic groups than the ACF code; however, both codes well complement each other. The ES vector code is based on the counting statistics of the number of routes in the coloured molecular graph mapped onto the coded compound. It provides a way for performing a detailed (though highly redundant) description; therefore the principal component analysis [28] of encoded data was applied in the next stage. To characterise each molecule, the contributions from the first eight principal components responsible for 99% of changes of complete code were used. Experiments have shown that the structure encoding system [148] is rather well suited in this case.

Various procedures for estimating the degree of structural similarity have been considered.[151] In particular, similarity of structures X and Y characterised by vectors $\mathbf{X} = \{x_1, \ldots x_i, \ldots, x_n\}$ and $\mathbf{Y} = \{y_1, \ldots, y_i, \ldots, y_n\}$ can be described using the Tanimoto coefficient

$$S_T = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \sum_i y_i^2 - \sum_i x_i y_i}.$$

The **X** and **Y** vectors can be composed of zeros and unities representing the presence or absence of the fragment, respectively, or of integers representing the occurrence number of the fragment in the structure. As applied to IR spectra, the 0/1 representation is more informative than the integer representation, since the absence of any fragment often changes the entire IR spectrum, whereas the difference in the number of the same structural

fragments usually results in a gradual change in the spectral pattern. Therefore, the structures to be sorted are selected using 0,1 vectors and ranked in accordance with the similarity criterion calculated on the basis of the occurrence number. Ranking of molecules using ES vectors is based on calculations of Euclidean distances from the vector of the target molecule.

Experiments were carried out using the IR spectra recorded in the region 2250 – 550 cm^{-1} with a resolution of 2 cm^{-1} and represented as sets of 128 points separated by 13.5 cm^{-1} (this corresponded to the same spectral resolution); at each of the 128 points, the value was averaged over seven points in the experimental spectrum covering the 13.5 cm^{-1} interval. Spectral intensities were normalised so that the absorbance of the most intense band was taken as 1.0.

Reliable estimate of similarity in the IR spectra is by no means easy problem. The spectral pattern seen by a spectroscopist's eye is hard to digitise. To solve this problem, the Pearson rank correlation coefficient and Spearmen rank correlation coefficient (the strongest, medium intensity and weakest peaks are respectively characterised by the highest, mean and lowest ranks) were used.[148]

The spectrum prediction algorithm was as follows.

1. Search for the nearest neighbours in the structure space. Compare the target structure with all reference structures in the collection and select k nearest neighbours for each of three models of structural similarity (the Tanimoto coefficients for both the 0,1 and integer vectors and the Euclidean distances for ES vectors). Get three lists of coincidences, containing information on k structures each.

2. Rank the selected molecules in accordance with the three measures of similarity. Search for the same structures in the three lists and sum up the individual ranks. If some molecules were found in one or two lists only, then ascribe rank $k + 1$ to these molecules in those lists in which they were not found. Sort molecules in ascending order of complete ranks and generate the final list.

3. Select those molecules whose characteristics either exceed the structural similarity threshold or are below the distance threshold.

4. If the final list contains information on less than two molecules, then terminate the procedure and display the results. Go to step 6.

5. Check the IR spectra of selected molecules for spectral homogeneity using correlation coefficients. If more than 50% of the spectra (*e.g.*, 2 out of 2, 2 out of 3, 3 out of 4, *etc.*) are much alike (*i.e.*, the correlation coefficient value exceeds a certain threshold value), then use the averaged spectrum for prediction. If the checking for homogeneity fails, then return control to the chemist. Go to step 6. In any case, display all available information.

6. Reject automated prediction since no or only one molecule is available. Prompt the chemist to assess the situation and decide between continuation and termination of the program. If a decision to predict the spectrum is made, then prompt the chemist to specify the molecules that were ascribed the highest rank and whose spectra should be used.

Thus, prediction can be performed either automatically or manually. If the chemist makes final decision, the result of prediction is characterised by higher quality. The CPU time is dependent on the number of fragments in the structure and takes several seconds on a personal computer with a Pentium CPU operated at 75 MHz, which is well suited.

Unfortunately, the authors restricted themselves to demonstration of a unique example (Fig. 5), which shows that the predicted and experimental spectra are similar. However, no attempts were made to test the method by sorting actual answer files of the ES, which makes impossible to obtain a statistically justified estimate. In our opinion, this approach has the advantage that its algorithm is free from the 'black box' type elements; therefore, the spectroscopist can easily establish the reasons for change in the prediction quality.

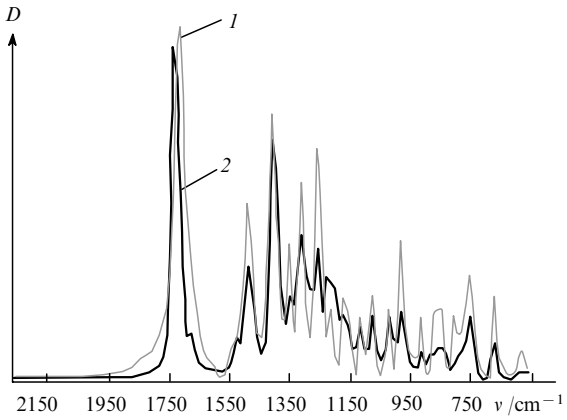


Figure 5. IR spectra: experimental (*1*) and predicted on the basis of search for the nearest neighbours of the target molecule (*2*).

V. Pathways of development of expert systems

Among the expert systems created in 1970 – 1980s, the best known is the DENDRAL system[8] referred to as a 'classical' example of ES in numerous publications on chemistry and books on AI. Popularity of the DENDRAL system is, on the one hand, to a great extent associated with the fact that this project was worked out under supervision of the Nobel Prize winner J Lederberg and noted scientists in the field of AI. On the other hand, the authors of the project have reported about 40 articles collectively called 'Application of artificial intelligence in chemistry', which could not but drawn the attention of many specialists. However, Gray[152] critically analysed the idea behind the DENDRAL system and concluded that this artificial intelligence system is far from being perfect. This problem was hotly debated[153, 154] with participation of the directors of the project. In contrast to a widespread opinion that the CONGEN program has found extensive application, Gray stated that it used rarely (he knows 8 articles in which application of the system was reported). One of the reasons for not-too-high popularity of the system is probably that it is not user-friendly.

In this review the emphasis is placed on the modern ES appeared in the 1990s. The potentialities of modern ES and those belonging to earlier generations were compared.[155] In the last decade, the efforts of researchers were aimed to improve the methods for retrieval of structural information from the spectra, to search for optimum strategies of computer-assisted structure recognition and to increase the user-friendliness of the systems. Our experience has shown that non-user-friendly ES have no chances to become an analytical tool for the chemist. A user-friendly system must:

(1) have convenient and operationally simple tools for information interchange with the user, which allows the chemist control over any step of the solution of the problem;

(2) offer the chemist the possibility of getting explanation of each decision made by the program;

(3) inform the chemist in case of 'emergency' at any solution step and display messages adequate to the situation;

(4) have advanced graphical tools for representing spectral and structural information to allow a chemist to quickly assess current status of the program operation and make corresponding decision.

In this review we consider certain ES, which, in our opinion, reflect modern tendencies and new approaches. However, let us first define basic concepts to be used later. Let the empirical formula of an unknown compound be determined and the number of isomers corresponding to the empirical formula be N. In principle, the N value can be found by structure generation using appropriate software. We say that a set of n structures ($1 \leqslant n \leqslant N$)

selected after the ES operation is a solution of the problem. The solution is correct if it contains a proper structure, otherwise it is incorrect. A correct solution and proper structure can be found only by checking each structure from the answer file using prediction of its spectra and additional experiments. If the answer file is empty ($n = 0$), the problem has no solution consistent with experimental data, information stored in the knowledge base and system options.

A description of the RASTR expert system created to elucidate the molecular structure using IR and NMR spectra is available (see, *e.g.*, Ref. 22). This system is capable not only of revealing possible structures, but also of generating, constructing and depicting spatial models of all stereoisomers corresponding to a given structural formula. The RASTR system was used as a prototype and basis for a user-friendly X-PERT system [29] which operates under MS WINDOWS. Let us consider the most important elements of this ES, which in one form or another are incorporated into several other systems.

1. The X-PERT system

The X-PERT system was created to elucidate the structure of organic molecules containing up to 30–35 skeletal atoms (C, N, O, P, S, B, Si, Hal) using IR, ^{1}H and ^{13}C NMR spectra. The knowledge base of the system comprises a set of libraries of molecular fragments and ranges of changes in their characteristic spectral features. The libraries are composed taking into account that the fragments belong to different classes of chemical compounds and respond specifically toward different types of spectroscopic techniques. Libraries belonging to different categories play different roles. The first category comprises universal libraries used for both fragment selection and structure filtration. Libraries belonging to the second category are adjusted to different types of spectroscopic techniques and are mainly used as filters for structure verification. Each fragment in every library is associated with its characteristic features in IR and NMR spectra, which are also stored in the knowledge base.

There are four hierarchically organised universal libraries.

1. Library of principal functional groups (LPF), *e.g.*, C=O, C≡N, *etc*. The widest ranges of changes in characteristic spectral features are used for these groups irrespective of the nearest environment of a fragment. These libraries are mainly used for preliminary evaluation of a specimen to assign it to a particular class of chemical compounds and for recognition of the structure of non-additive compounds.

2. Library for structural-group analysis contains functional groups and information on the admissible nearest environment, *e.g.*, $-\overset{|}{\underset{|}{C}}-CO-\overset{|}{\underset{|}{C}}-$, $-\overset{|}{\underset{|}{C}}-O-\overset{|}{\underset{|}{C}}=O$, $-CH_2-OH$, *etc*. The library is used for SGA and structure filtration.

3. Library of aromatic fragments (AROM) contains 12 types of benzene ring substitution.

4. Library of phosphorus-containing fragments.

There are the following specialised libraries:

A library of ^{13}C NMR data; three libraries of ^{1}H NMR data to reveal CH_3, CH_2 and CH groups with consideration of the nearest environment; a library containing fragments (mostly cyclic) characterised by strong specific absorption bands (*e.g.*, four- and five-membered cyclic ketones, lactones, lactams, *etc*.); and a library of fragments containing an OH group.

For each fragment, the system generates a 'visit card' containing the following information: the fragment structure; ranges of changes in the characteristic features in IR (frequencies, halfwidths and intensities of absorption bands) and NMR spectra (chemical shifts, SSCC, signal multiplicities and corresponding numbers of C or H atoms); structural descriptors used to set the requirements for the fragment environment (additivity conditions) and provide retention of its spectral features on going from one molecule to another. Descriptors characterise the fragment atoms with free valence bonds. The descriptor indicates the possibility of adding a hydrogen atom, hetero-atoms (they can be enumerated), aromatic cycles and multiple bonds to a given atom. The possibility (or impossibility) of the formation of a multiple bond and incorporation of the fragment into cyclic structures of specified size is also indicated.

The descriptor can have three types of meanings: 'possible' (P), 'forbidden' (F) and 'obligatory' (O). The program finds classes of topologically equivalent atoms with free bonds in the fragment structure and colours them using individual colours. Each coloured atom is associated with a column of corresponding colour in the table of descriptors. This makes it possible easily to form, interpret and correct the values of descriptors characterising equivalent atoms. The total number of fragments in the libraries is 450.

Experimental data (the empirical formula or molecular weight and available spectra) can be input manually or imported using the OPUS and WIN-NMR programs. The system checks the data for consistency and prompts the chemist to double-check the reliability of the data in questionable situations. If the empirical formula is unknown or questionable, the program generates all possible molecular formulae corresponding to the specified molecular weight taking into account the constraints imposed on the qualitative and quantitative composition of the substance. In this case the most probable molecular formula is chosen using a recently developed method.[156 †]

From this time on, the system operates in the interactive mode. This makes it possible to use the experience and intuition of the chemist and the capability of computer to perform fast search for necessary information and complex logic-combinatorial calculations. In this case, the basic requirement for ES is also met: at any step in the solution the program can explain the reasons for one or another decision and the user can be prompted to activate a corresponding option.

Prior to starting the operation, the program assesses complexity of the problem and displays messages concerning the method of its solution. If the problem can be solved automatically, the program generates all structural isomers and selects possible structures by passing them through a system of filters.

The first stage of the standard method for solution is generation of fragment sets. First, possible fragments are selected with consideration for their empirical formulae, spectral intervals and the unsaturation degree. In this case, necessary libraries of fragments are chosen by the user. The fragments are selected taking into account that the reliability of finding a fragment is 100% if the experimental band falls into the middle of the characteristic interval. It (reliability) linearly decreases as the interval bounds are approached and decreases to zero on both sides of the bounds. A trapezoidal membership function [19] of the interval is chosen so that the decrease in the reliability near the bounds with increasing the length of the interval be more gradual.

The fragments selected and their reliability indices are displayed. The 'Show Spectra' procedure makes it possible to represent experimental spectra and characteristic intervals for selected fragments in clear graphical form to allow the chemist to assess the reliability of the fragments and to edit the list by adding (if needed) his own fragments. Then the table with interpretation data of all available experimental spectra is displayed. The chemist can analyse logical relations between the fragments and correct the 'reasoning' of the program using his experience and *a priori* information. The final set of consistent logical equations is used as input information for the 'inference machine'. The result of the solution of the logical equations [function f(A), see Section II.1] is displayed as a set or several sets of fragments consistent with the spectra, empirical formula and the unsaturation degree of the compound analysed. The program automatically adds sets of

† See also a recent study by M E Elyashberg, Yu Z Karasev and E R Martirosian *Anal. Chim. Acta* **388** 353 (1999).

microfragments such as CH_3, CH_2, *etc*., which were not included as constituents of the structures of selected fragments. The chemist can also correct the selected fragment sets or compose his own sets. The special-purpose 'Multiple Fragments' procedure allows for the inclusion of a fragment appearing repeatedly in the structure, which reduces dimensionality of the problem.

Structure generation using fragment sets is the next step of the program operation. In this stage, the possibility of imposing structural constraints listed in Section III (GOODLIST, BADLIST, constraints associated with the ring size, bond multiplicities, *etc*.) is provided. There is an additional option which enables or cancels the descriptor control in the course of structure generation. The descriptor control is useful to preclude generation of a large number of wrong (bad) structures. Cancellation of the descriptor control can appear to be useful in the case of repeated solution of the problem if the structures found are questionable. The structure generation is carried out by the program based on the known algorithms.[84,85] The structure counters allow the chemist to monitor the process and assess the efficiency of the constraints imposed.

Structure generation using fragment sets and free atoms can lead to such substructures whose presence contradicts the spectra and/or concepts of organic chemistry, therefore all generated structures are verified using the following six filters (see Section IV).

1. Checking for consistency with characteristic spectral features.
2. Rough simulation of ^{13}C NMR spectra (see Section IV.1.a).
3. Checking for consistency with principles of structural chemistry and stereochemistry (unlikely fragments, Bredt's rule).
4. Checking each isomer for its appropriateness to be used in interpreting experimental spectra.
5. Search for and exclusion of identical structures.
6. Prediction of chemical shifts in ^{13}C NMR spectra to reveal the most probable structures.

The system chooses necessary libraries on the basis of generalised 'portrait' of the answer file to perform a two-stage check. In the first stage, the list of fragments that contradict the spectra is generated. In the second stage, only contradictory fragments are sifted using structural formulae to save time and reject those structures in which at least one of these fragments is found. Options used in the filtration stage allow the user to specify the parameters determining hardness of constraints imposed (*i.e*., the degree of fuzziness of the interval bounds).

Structure verification using the first and second filters can be performed in both the standard and supervised modes. The first of them provides automated filtration (the so-called 'blind' filtration). In this case, only messages on the current status of program operation and the running number of 'good' and 'bad' structures are displayed. In the second mode, the filtration process is visualised and the user can control over it as well as get all information necessary to explain the decisions made by the system. Each structural formula to be checked is displayed. If no contradictory fragments are found in the structure (it should be kept in mind that only these are sifted), the structure is indicated as good. Otherwise contradictory fragments found in the structural formula are coloured and information on characteristic spectral features (frequencies, chemical shifts, parameters of bands or signals) not confirmed by the spectra is displayed in the form of a 'visit card' of the fragment.

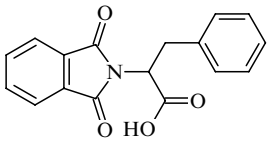
Depending on the number and type of contradictory fragments found in the structure, the user decides to reject the structure as unlikely or accept it for further checking. Structure filtration using this user-friendly mode of the program operation is visualised, fully controllable and understandable.

Checking the number of signals with different multiplicity in ^{13}C NMR spectra in the supervised mode is accompanied by displaying the data on the differences between the number of predicted and observed signals.

After passing the data through the first five filters the system automatically uses the answer file as input data for the SpekEdit program which predicts the chemical shifts in ^{13}C NMR spectra. Then the standard deviations (S) of calculated spectra from experimental ones are calculated and the structures are ranked in ascending order of S.

The potentialities of the system will be assessed below; here we consider only one example to illustrate the operation of the program.[29]

Let an unknown compound have a molecular weight of 295 and possible composition $C_{1-50}\,H_{0-120}\,N_{0-30}\,O_{0-30}$; 197 different molecular formulae correspond to this set of parameters. Using IR and NMR spectral data, the program found the only molecular formula $C_{17}H_{13}N_1O_4$ and created the fragment sets (1) C=O, C_6H_5, 1,2-C_6H_4 and (2) C=O, C_6H_5, 1,3-C_6H_4. These sets were used to generate 6995 structures. The proper structure



was elucidated using spectral and structural filtration.

2. The SpecSolv system

The SpecSolv system ‡ is a new module of the well known SpecInfo system [157] designed for storage and structural interpretation of NMR, IR and mass spectra. In particular, the SpecInfo knowledge base consists of over 200 000 assigned ^{13}C NMR spectra which serve as a basis for the development of the program.

The SpecSolv system is designed to elucidate the molecular structure using the ^{13}C NMR spectrum only and does not require the knowledge of the molecular weight and molecular formula. It uses parameters of ^{13}C NMR spectra [the chemical shifts, signal multiplicities and intensities (the latter as the number of carbon atoms)] as input data. The knowledge base of the system is a library of fragments and their subspectra. The library was created by automated 'cutting off' fragments from the structures stored in the SpecInfo database and contains more than 400 000 fragments encoded by the three-sphere HOSE code [73] and more than 100 000 fragments encoded by the two-sphere HOSE code with respect to the central carbon atom. The library is carefully checked by experts.

Information on a fragment includes the connectivity matrix, HOSE codes of all heavy atoms and the subspectrum. Parameters of ^{13}C NMR spectrum (chemical shift, multiplicity, intensity and root-mean-square deviation R of the chemical shift) are stored for each carbon atom. The R values were calculated for the sets of equivalent substructures in the SpecInfo database. All fragments are completely defined including ring closures in the outer sphere.

The first step of structure elucidation with the program is the fragment search. All subspectra in the library of fragments are compared with the experimental spectrum and the fragments whose subspectra match corresponding regions in the experimental spectrum are selected. A typical case is selection of $\sim$ 500 fragments for a molecule containing $\sim$ 25 heavy atoms.

Fragment search includes three stages: preliminary (presearch), main and fine search. After the first stage the subspectra are coded as strings encoding the presence of signals with a certain multiplicity within discrete regions of the ^{13}C NMR spectrum. At the second stage it is tested as to whether the number of signals in the regions selected after presearch has a matching number of lines in the experimental spectrum. The fine search includes checking for exact deviations of the chemical shifts in both spectra.

‡ The first article contained the description of the SpecSolv system [30] appeared in 1996 and had an intriguing title 'Fully Automated Structure Elucidation — A Spectroscopist's Dream Comes True'.

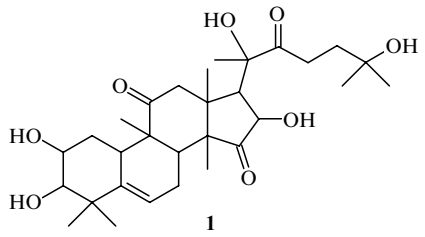
Typically, the deviation is 2 to 5 ppm. The selected fragments are then ranked in descending order of match factor, which is defined as the mean deviation of the chemical shift of each atom from the signals in the experimental spectrum.

Despite the fact that the program can operate without knowledge of the molecular formula, it is necessary to specify possible types of atoms (especially, hetero-atoms) since they cannot be immediately revealed by ^{13}C NMR spectroscopy.

It is essential that there is no need for a chemist to analyse all selected fragments, since bad fragments should be excluded at the stage of structure generation. However, if *a priori* information is available (most often, such is indeed the case), it can be expected that analysing the list will make it possible to exclude the bad fragments to save the working time. The main feature of the generation algorithm used in the SpecSolv system is that the fragments are linked by overlapping common atoms. The largest fragment whose subspectrum best matches a particular region of the experimental spectrum is taken as the first one and the program tries to add successively any other fragment to this fragment. If the new fragment passes the plausibility test, it is accepted and the enlarged fragment is considered as starting structure. Before adding the next fragment, the program predicts the ^{13}C NMR spectrum of the enlarged fragment. This is achieved by simple combination of the subspectra, which provides a high performance of prediction. Intermediate structures for which the deviations of the chemical shifts exceed a prescribed value are excluded. In this fashion the final structure is being assembled at the same time with cross-validation of intermediate structures. The absence of such a verification tool in structure generators that admit assembly of structures from overlapping fragments (*e.g.*, GENOA [87]) is a grave drawback.

The assembly continues until all chemical shifts in the experimental spectrum are assigned and free bonds in the resulting structure are absent. Usually, 10 generation cycles based on the first 10 best (the largest with minimum deviation of the chemical shifts) fragments are sufficient to elucidate the answer structure. This is possible since the selected fragments are ranked by match factor of their subspectra with the experimental spectrum. Most often, only one structure is found after exhaustive search in the entire fragment space, though several isomers of similar structure can also be obtained (in general, if no empirical formula is given, generation of nonisomeric structures is also possible).

Will, Fachinger and Richert [30] reported recognition of rather large molecules, *e.g.*, such as **1**.



1

However, it is obvious that structure generation is possible only if the fragment library contains large overlapping fragments of such molecules. Hence successful identification requires the presence of a rather large number of large molecules related to the molecule to be recognised in the system database. However, even if the database contains appropriate fragments, the absence of common atoms will make generation of the proper structure impossible, which means a strong dependence of the result on the diversity and quality of the database. The program is also very 'sensitive' to the solvent in which the experimental spectrum is recorded and, in addition (this is a grave drawback) does not recognise stereoisomeric forms. These factors can have a determining effect on the results of comparison of predicted subspectra of intermediate substructures and the overall model spectrum with the experimental one. It is reported that the system elucidated the structure of 80% of all compounds analysed in the laboratory. The system was adapted for exploitation on a VAX 6610 minicomputer.

As can be seen, at present 'a spectroscopist's dream' does not come true with the aid of the SpecSolv system since 20% of all problems remain unsolved. If the system cannot solve the problem, a corresponding message is generated and the program is terminated. Therefore new approaches are required which make it possible to resume the analysis in those cases where the program reaches a deadlock. A procedure for overcoming a deadlock has been implemented in the ACD/Structure Elucidator system.[31]

3. The ACD/Structure Elucidator system

The ACD/Structure Elucidator system integrates the recognition strategy developed by the authors of the SpecSolv system [30] and conventional methodology used in the RASTR,[22] X-PERT,[29] CHEMICS [37] and other systems.

Two structure elucidation modes, a standard mode and the classical ES mode are permitted. The first mode is to a great extent based on the known general strategy.[30] As in the SpecSolv system, the set of experimental data includes a ^{13}C NMR spectrum, chemical shifts, signal multiplicities and intensities. If it is difficult to indicate the exact number of carbon atoms assigned to a signal, their minimum and maximum values are specified. To facilitate measurements of signal intensities, it is appropriate to record a ^{13}C NMR spectrum with an admixture of a paramagnetic reagent, *e.g.*, $Cr(acac)_2$, which shortens and equalises the spin-lattice relaxation time T_1 for all carbon atoms including quaternary ones. The experimental spectrum can be imported directly from the spectrometer or the user can input spectral parameters manually. Usually, only an NMR spectrum is required for structure elucidation; however, it is helpful to have IR and ^{1}H NMR spectral data to facilitate the solution of the problem. The database of the system is a library consisting of more than 400 000 fragments and their ^{13}C NMR subspectra that were derived from 135 000 structures with assigned ^{13}C NMR spectra (the structural database). This ES has an appreciably smaller knowledge base than the SpecSolv system has.

The operation of the program begins with the search for ^{13}C NMR spectrum in the knowledge base. Once a spectrum is found, a corresponding structure is displayed and the program is completed. Otherwise the structure is reconstructed from fragments.

At first, the program selects fragments whose subspectra do not contradict the experimental spectrum. The fragments found are sorted in descending order of the number of skeletal atoms and mean deviations of their subspectra from corresponding signals in the experimental spectrum. Since it is impossible to accept or reject the presence of, *e.g.*, OH, NH and $C \equiv N$ groups or to differentiate between δ_C of the $O=C-N$ group and aromatic ring signal (150–160 ppm) using the ^{13}C NMR spectrum, then, if the IR spectrum is available, filtration of selected substructures is performed using a library of functional groups having the most characteristic features in the IR spectra. Analogous filtration can be performed using fragments having clearly defined characteristic features in the ^{1}H NMR spectra. After filtration the number of selected fragments is reduced (in some instances, it becomes halved), which favours placing good fragments at the beginning of the list. Then, beginning with the first fragment, the fragments are checked for the possibility of linking by common atoms to perform successive assembly of the structure. As in the SpecSolv system, linking the fragments is accompanied by prediction of the ^{13}C NMR subspectrum of the enlarged fragment to verify the correctness of the assembly.

The problem is considered solved if after checking of all possible combinations of the fragments the program generates one or several structures. The most promising fragments (*i.e.*, the largest ones whose subspectra match well the experimental spectrum) are placed at the beginning of the list; therefore the answer structure is often generated within minutes after starting

the operation of the program. In the case of good correspondence between the ^{13}C NMR spectrum predicted for the answer structure and the experimental spectrum the process can be terminated before completion of exhaustive search for all combinations of the fragments.

If the attempt to assemble the structure fails (the knowledge base contains no appropriate fragments having common atoms or good fragments were not selected), the program automatically switches on the classical ES mode to operate with DUS.

As was mentioned in Section III, operation in this mode requires knowledge of the molecular formula or at least the molecular weight. If the molecular formula is unknown, the program first selects functional groups that would be present in the molecule analysed. To this end, a library containing a small number of the most characteristic functional groups and their spectral features is used. Then search for selected functional groups in the structure of previously selected fragments is performed. The elemental composition and unsaturation degrees of the verified functional groups, as well as the numbers of carbon and hydrogen atoms counted from NMR spectral data are used by the program for automated determination of structural constraints for the generator of empirical formulae. One or several possible empirical formulae established by this generator are passed to the structure generator for verification.

This structure generator operates with the atoms and fragments having no common atoms. It is based on the known algorithms.[84–86] The fragment sets are composed as follows.

The first N (reference) fragments in the primary list of selected and filtered fragments (usually, the chemist varies the N value from 10 to 30 on his own) are chosen. Obviously, these are the largest fragments and this set almost always includes those fragments that are constituents of the structural formula of analysed compound. Then, the fact is used that, according to the ranking algorithm, small fragments (most often functional groups with their nearest environment) are placed at the end of the primary list.

The subspectrum of the first reference fragment is compared with the experimental spectrum to select the unmatched chemical shifts in the latter. Then, small fragments appropriate for interpreting these chemical shifts are chosen. Only those combinations of small fragments are added to the first reference fragment that provide the highest degree of interpretation of the experimental spectrum. If thereafter certain signals in the query ^{13}C NMR spectrum remain uninterpreted, the program adds corresponding microfragments (CH_3, CH_2, CH, C=O, *etc.*) to the set. Each of the sets obtained is checked for consistency with the current empirical formula of the molecule. The sets corresponding to the second, third, *etc.* reference fragment are generated analogously. If no small fragments appropriate for composing the sets were found, the latter are formed using microfragments so that the number of sets becomes equal N.

The chemist can add his own fragments to the sets, compose the sets based on one of the reference fragments and use all types of constraints described in Section III. To facilitate this, the program displays a 'generalised portrait' of the molecule under study, *i.e.*, the distribution of typical functional groups over frequency of meeting them in the selected fragments. High frequency of meeting indicates that the group is, probably, a constituent of the molecule, whereas the absent groups are used to generate the BADLIST.

Before starting the structure generation using the current fragment set the program assesses the working time (to this end, the percentage of the job done in a time set using system options, *e.g.*, 20 s, is measured) and, if it appears to be longer than a reasonable time, this set is skipped during the first program run and the program begins testing the next fragment set. All structures resulting from the structure generation process are checked for consistency with the available spectra and principles of organic chemistry and stereochemistry. A special program developed to search for tautomeric structures excludes unstable forms from the list, which often substantially reduces the size of the answer file.

Then the structures automatically arrive at the input of the program for prediction of ^{13}C NMR spectra (the ACD/CNMR Predictor program[118]). When prediction is completed, the structures are ranked in ascending order of mean deviations of the predicted spectra from the experimental spectrum. Our experience shows that calculations of the spectrum using a personal computer equipped with a Pentium CPU operating at 133 MHz take from 0.5 to 1 s, so prediction of the spectra of 1000 structures takes only 5 to 15 min depending on the number of carbon atoms. Testing of the system showed that in the majority of cases the proper structure is ranked 1st or 2nd from the beginning of the answer file or is at least placed among the first dozen of the structures. However, if the difference between the chemical shift deviations of the preferred structure and the succeeding structure lies within the accuracy of the calculation of the ^{13}C NMR spectrum, further verification of the solution is required.

To this end, automated prediction of ^{1}H NMR spectra is performed using the ACD/HNMR Predictor program[137] (see Section IV.2). As a rule, only the spectra of the first dozen of structures are predicted, so the CPU working time is relatively short ($\sim 10-60$ s per structure depending on the complexity of the spin systems calculated). Experimental and calculated ^{1}H NMR spectra are compared by a chemist who makes the final choice of the preferred structure or makes a decision on the necessity of additional experiments. The calculated spectra, integral curve and δ_H values for all lines are displayed. If the predicted spectra of several structures appear to be equally close to the experimental spectrum, the proper choice may be made by using the Δ_{CH_3} values for the δ_{CH_3} singlets

$$\Delta_{CH_3} = |\delta_{CH_3\ exp} - \delta_{CH_3\ calc}|$$

and comparing the integrals for all multiplets. In most cases, at $\Delta_{CH_3} \geqslant 0.2-0.3$ ppm the structure can be rejected. Resolving of questionable situations requires additional experiments, in particular, using 2D NMR spectroscopy.

Thus, the ACD/Structure Elucidator system differs from other systems in the possibility of using two interacting strategies for solving the problem, each of them being based on corresponding structure generator. If in both operating modes the program cannot automatically find an adequate solution, the possibility is provided of solving the problem in the step-by-step mode under the supervision of a chemist who has considerable opportunities to interfere in the 'reasoning' of the program. Testing of the system on a test set of 220 problems showed that correct solution can be found in 90% of cases.

4. 2D NMR spectroscopy in expert systems

The systems created have progressively been developing. In addition to improving the algorithms and extending the knowledge bases, attempts are made to use 2D NMR spectroscopic data in ES, since methods employed in this area of NMR spectroscopy have received wide acceptance and, actually, became routine.

For instance, instrumental tools for using 2D NMR spectroscopy data were incorporated into the CHEMICS system to extend its capabilities.[158,159] Methods of processing the data obtained in 2D-INADEQUATE (C–C correlations through one bond) and $^{1}H-^{13}C$ NMR COSY (C–H correlations through one bond)/$^{1}H-^{1}H$ NMR COSY (H–H correlations through three bonds) spectroscopic experiments were developed. Using the data of 2D NMR experiments, it is possible to reveal CH_n groups ($n = 0-3$) bonded by chemical bonds and to correlate this additional information with the library fragments selected by the system, thus testing the possibility of the formation of a bond between the fragments. The data obtained are used for controlling the process of structure generation from the fragments.

A new SESAMI system[161] was created as a result of improving the CASE system.[160] The library of ^{13}C NMR spectra containing 5100 ACF with one-layer environment is used as the

database in this system. First, SESAMI generates an exhaustive list of fragments, which then is made more 'sparse' using a special-purpose subprogram. Another subroutine adds extra constraints imposed on the structures. The reduced list of fragments and the constraints are used as input data by the COCOA structure generator.[79] Preliminary selection of the structures generated is performed on the basis of prediction of their ^{13}C NMR spectra and an expert makes the final decision.

Enhancement of potentialities of the SESAMI system is attained by incorporating the software [162] developed for processing the data of 2D NMR experiments. The system receives information on correlations between signals and builds fragments consistent with these data. The structures are generated using the COCOA program. Further improvement of the system is performed to develop a module based on application of ANN and designed to elucidate the fragments in the case of simultaneous use of IR and ^{13}C NMR spectra.

In parallel with improving the known ES by incorporating tools for processing 2D NMR data, investigations on creation of computer systems designed to retrieve and efficiently use structural information 'hidden' in 2D NMR spectra for molecular structure generation began in the last decade. The demand for the development of such systems is caused by the necessity of establishing the structures of very large molecules of natural products containing up to 50–60 skeletal atoms. Of interest is a recently developed approach.[32, 163, 164] The authors started from a rather logical position that if once enormous efforts were made to isolate a natural product, carrying out a number of rather labour-consuming 2D NMR experiments seems to be entirely justified. Fragments designed immediately in the course of experiments are mostly used for structure elucidation. As a result, a CISOC-SES expert system was created whose operation requires the data of 1D (^1H and ^{13}C) and 2D DEPT, DQF-COSY, HMQC, HMBC, HETCOR, COLOC and 2D-INADEQUATE NMR experiments. In this approach, structural data obtained from 2D NMR experiments are mostly used (the structure is reconstructed as if by applying 'the first principles'), therefore the authors called this approach '*ab initio*'.

Obtaining information from 2D NMR experiments requires a skilful choice of procedures (pulse sequences) and correct use of data. Hence the efficiency of such systems will be to a great extent determined by the success in implementing the man–computer 'symbiosis'. Since the systems based on fragment selection retrieve valuable structural information, one should expect that both approaches will eventually integrate.

VI. Strategy of molecular structure elucidation

Since the requirements for the ES architecture are, on the whole, already formulated, the development of an optimum strategy of the ES application is topical. This problem can hardly be solved theoretically because it is impossible to imagine the whole versatility of situations that may arise in the course of structure elucidation. Efficient approaches can be found only after analysis of the solution protocols of a rather large number of problems.

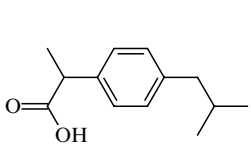
An attempt to develop the methodology of the molecular structure elucidation using ES was made.[141] Particular emphasis was placed on the study of difficulties and contradictions that arise when solving the problems and on the development of procedures for overcoming them. The authors carried out a large-scale computational experiment using the X-PERT system.[29] About 200 structural problems taken mostly from monographs [165–168] or spectral databases and analytical practice were solved. The molecules analysed contained from 10 to 23 skeletal atoms, which is characteristic of the compounds commonly used in industrial laboratories. Most of empirical formulae contained hetero-atoms (from 1 to 10). Over 80% of the total number of compounds belonged to the cyclic ones and about 45% of them contained benzene rings.

As a rule, experimental data included IR, ^1H and ^{13}C NMR spectra, as well as the empirical formula or the molecular weight of the substance. To find possible ways of solving the problems, each of them was repeatedly solved using different methods and strategies. Particular emphasis was placed on elucidation of the reasons for difficulties and failures. To facilitate the logic-statistical analysis of the data obtained, the system automatically stored the solution protocols as text files incorporated into the active database. Plausible structures were elucidated on the basis of prediction of chemical shifts in the ^{13}C NMR spectra using specially developed software.[169] Computational experiments were carried out on an AT 486 personal computer operated at 66 MHz.

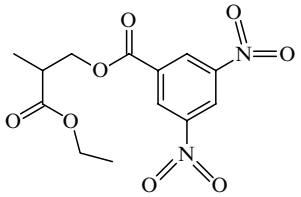
Analysis of the protocols has shown that it is possible to find a correct solution in virtually all the cases considered using the interactive mode. The percentage of unambiguous solutions was 40% and the number n of candidate structures did not exceed 5 for 70% of the problems. The n value exceeded 10 for 15% of the problems. Prediction of ^{13}C NMR spectra of ambiguously solved structures increased the total percentage of unambiguous and right answers from 40% to 75%, *i.e.*, the proper structure was ranked 1st and placed at the beginning of the file.

For instance, the answer structures of molecules **2**–**9** were obtained with indication of n values, reduced size of the answer file after prediction of ^{13}C NMR spectrum (n^p) and the number of the proper structure in the ranked answer file n^t. The n^p value was defined as the number of structures for which the root-mean-square deviation s does not exceed 5.5 ppm (the accuracy of the program [169]).

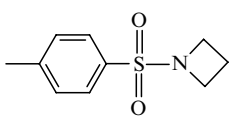
About 60% of problems were solved automatically by using SGA or straightforward structure generation (SSG) from empirical formulae. In those cases where the structure was unambiguously elucidated by means of SSG followed by structure filtration, exhaustive information on the molecular structure was obtained



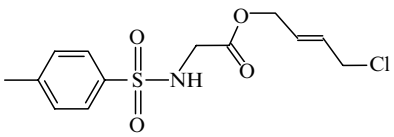
2 ($n = 5$, $n^p = 4$)



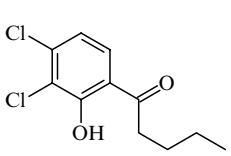
3 ($n = 2$, $n^p = 1$)



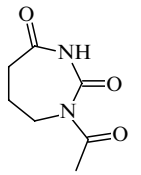
4 ($n = 4$, $n^p = 1$)



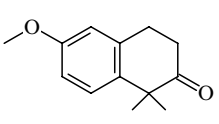
5 ($n = 7$, $n^p = 3$)



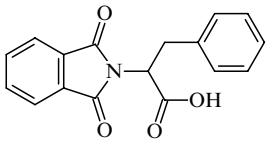
6 ($n = 6$, $n^p = 5$)



7 ($n = 4$, $n^p = 2$)



8 ($n = 4$, $n^p = 2$)



9 ($n = 4$, $n^p = 1$)

using characteristic features not observed in the experimental spectra. Moreover, filtration has led to recognition of certain structures containing fragments that were absent in the database.

In essence, the strategy of solving a problem is a set of methods for overcoming various difficulties. The authors [141] analysed typical difficulties encountered when using ES.

1. Examples of difficult situations

Mixed-valence atoms. Depending on the valence state, mixed-valence chemical elements can form various functional groups (*e.g.*, SH, $\rangle S=O$, $\rangle SO_2$, $\rangle PH$, $\rightarrow P=O$). Each functional group generates its own family of isomers. Therefore, strictly speaking, structural problems should be repeatedly resolved until all valence states of atoms present in the empirical formula have been searched through. The problem is appreciably simplified if available information on the origin of a specimen or data of independent experiments make it possible to determine the valences of all elements in the empirical formula.

Specific character of IR spectrum. It is known that IR spectrum is extremely sensitive to slight changes in the molecular structure and electron density. This makes it possible to reveal even minor changes in the molecular structure. However, frequencies or intensities can be interpreted only if the structural formula of the molecule is known. This is impossible until the molecular structure has been recognised. For this reason, rearrangements of functional groups in the isomers, which have little effect on the additivity of the molecule, can cause the characteristic frequencies and band intensities to go beyond the limits indicated in the fragment libraries.

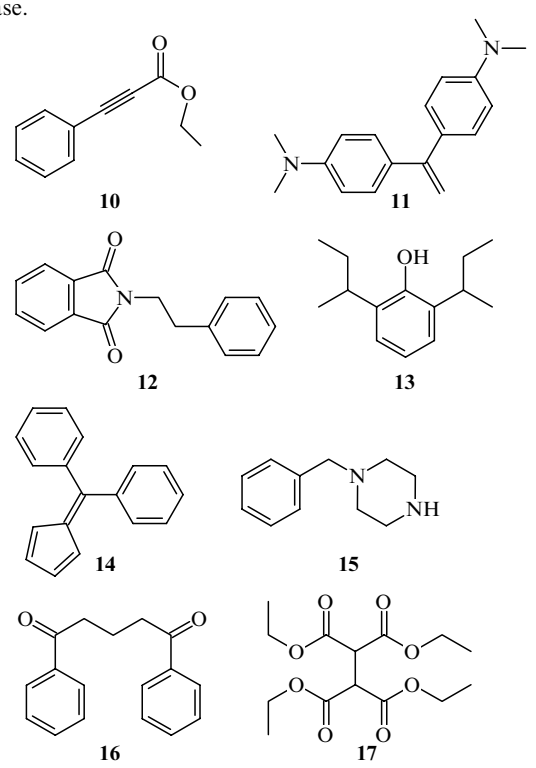
For instance, in some instances the bands at 3000, 1600 and 1500 cm^{-1} characteristic of benzene derivatives appear to be inactive or too weak to be detected in the spectrum. However, if the NMR spectra of such a compound contain signals in the aromatic region, the X-PERT system assesses the situation and prompts the chemist to accept or reject the presence of an aromatic fragment (in other words, to decide between 'yes' and 'no', respectively). Another problem arises if a compound is dissolved in a solvent of the CCl_4 or $CHCl_3$ type. In this case diagnostically important aromatic bands in the region 700–800 cm^{-1} cannot be used for determination of the types of benzene substitution, which results in increasing the size of the answer file.

The IR spectral pattern in the region of stretching vibrations of the OH group is dependent on many factors (the presence of neighbouring atoms and functional groups, solvent properties, *etc.*). Because of a very strong effect of the intramolecular $O-H\cdots X$ hydrogen bond, the parameters of some IR bands [ν_s(OH), ν_s(C=O), *etc.*] are shifted beyond the limits of characteristic intervals. Therefore, deciding between the presence and absence of the OH group often requires the interference of a chemist (the X-PERT system generates a corresponding message).

It turned out that joint use of IR and NMR spectra often leads to the loss of right fragments whose characteristic features were not confirmed because of unusual IR spectral pattern. Therefore, the IR spectrum should sometimes be 'switched off' during the first program session. Of course, it does not mean that the ES can exclude IR spectra from its 'reasoning', since there is a rather large number of functional groups which contain no C and H atoms (*e.g.*, NO_2, SO, SO_2, NO, ONO, P=O, P=S, *etc.*) and cannot be immediately detected by NMR technique without using the IR spectral bands for confirmation.

The influence of structural peculiarities on the complexity of the problem. At first glance, it may appear that the larger the molecule, the larger the expected *n* value. Actually, the *n* value is a very complex function of different structural parameters. Therefore it is quite possible that a large molecule will be identified unambiguously, whereas the answer file created when recognising a relatively small molecule will contain a large number of structures. Let us consider the most important structural factors affecting the problem solution.

If the symmetry of the structure of analysed compound is higher than C_1, the solution of the problem becomes easier, since a large number of wrong structures are often rejected after comparing the predicted number of signals and their multiplicities in the ^{13}C NMR spectrum with the experimental data. For instance, unique and proper answer structures were found for molecules **10**–**17**. It should be noted that structure **12** containing the O=C–N–C=O fragment was unambiguously recognised despite the fact of the absence of this fragment in the knowledge base.



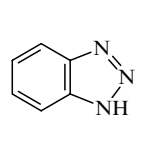
Symmetry of the structure also allows the chemist to use more efficiently the 'Multiple Fragments' procedure [29] to generate sets containing admissible (from the viewpoint of the empirical formula) numbers of identical fragments, which makes possible generation of large structures. For instance, a set containing four ester groups and four ethyl groups was created for the empirical formula $C_{14}H_{22}O_8$ and the unique structure **17** was recognised after the filtration procedure.

As is known, the number of signals in a ^{13}C NMR spectrum calculated as the number of groups of topologically equivalent carbon atoms does not always correspond to experimental data. For instance, only three signals are observed in the ^{13}C NMR spectrum of molecule **18** instead of the six expected. In this case the proper structure was unambiguously elucidated using prediction of ^{13}C NMR spectrum.

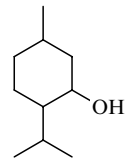
Accidental degeneracy of signals in the NMR spectrum or splitting of signals of carbon atoms of *gem*-dimethyl, isopropyl and other groups near the stereocentres makes computer-assisted structure recognition more difficult and requires interference by a chemist. If there are several stereocentres in the molecule, the answer file often contains a larger number of structures than usually. This is explained by complexity of ^1H NMR spectra (a small number or even the absence of the first-order signals, different δ_H values of geminal protons) and additional splittings of signals in the ^{13}C NMR spectra. For instance, in the case of compound **19** having three stereocentres the program for δ_C prediction ranked the proper structure 3rd from the beginning of the answer file containing 31 structures in the structural similarity criterion.

Large number of candidate structures in the answer file is also characteristic of hetero- and carbocyclic compounds, *e.g.*, for **18**–**25**. This is associated with a large number of rearrangements

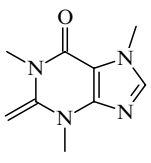
possible in heterocyclic structures with several substituents. Compound **20** can serve as an example. The set including all fragments of this structure, 3(>N−CH3), 2(>C=O), >C=C<, >C=N−, was created manually. The program generated 181 structures which were subjected to filtration to select 40 structures containing five- and six-membered rings.



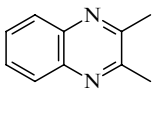
18 (n = **7,** n^t = 1)



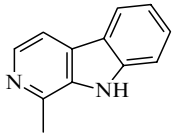
19 (n = 31, n^t = 3)



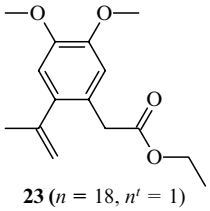
20 (n = 40, n^t = 14)



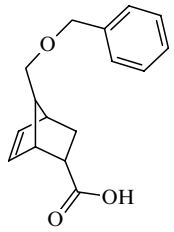
21 (n = 14, n^t = 1)



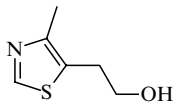
22 (n = 28, n^t = 4)



23 (n = 18, n^t = 1)



24 (n = 25, n^t = 1)



25 (n = 35, n^t = 1)

The S value of 11.6 ppm was obtained for the predicted ^{13}C NMR spectrum of the proper structure **20** (the maximum deviation was 25.6 ppm.). As a result, the proper structure was ranked 14th from the beginning of the list of structures in the structural similarity criterion. Thus, we get a 'fuzzy' solution even if the fragment set used is *a priori* suitable for recognition of the proper structure; therefore, elucidation of a proper structure requires additional information (obtained, *e.g.*, from 2D NMR experiments) and application of more reliable methods for spectra prediction.

Analysis of the ways of overcoming difficult situations made it possible to propose methodology resulting in correct solutions of the problems.

2. Strategies of problem solution

To obtain a correct solution of a problem using an ES, *i.e.*, to find a proper structure in the answer file containing a minimum number of possible structures (ideally, one structure), it is necessary (but not sufficient) to provide the possibility of selecting as many fragments constituting the analysed structure as possible. To this end, a large knowledge base should be incorporated into the system and characteristic spectral ranges of the fragments should be rather wide lest the correct fragments be lost.

When these conditions are met, it appears that the number of selected fragments is too large since many superfluous (wrong) fragments are also selected. This requires finding and excluding superfluous fragments. Editing the list of selected fragments is hardly possible without participation of a chemist. This procedure can be excluded only if all logically admissible fragment sets and generation of structural formulae are created automatically. As a result, the number of generated structures will be too large, structure generation and filtration will take a long time and the problem even may appear to be virtually unsolvable. In addition, the answer file can contain the structures that contradict *a priori* information. Hence, in any case checking the list of selected fragments by a chemist seems to be virtually unavoidable rather than desirable.

Thus, meeting the requirement for the presence of the largest possible number of right fragments in the list of selected substructures results in increasing the total number of fragments and, as a consequence, in increasing the size of the answer file. In this case the solution will likely be right, but 'fuzzy'. This should be considered as a pay for correctness.

The selectivity of fragment selection is strongly dependent on the environment descriptors of the fragments. The stiffer the requirements for the fragment environment, the narrower the intervals of characteristic spectral features and, hence, the higher the SGA selectivity. At the same time, the narrowing of the intervals increases the probability of the loss of correct fragments because of the deviation of an experimental feature beyond the bounds of the characteristic interval.

There is also a clear dependence between the size of the structural file and the rigour of definition of descriptors describing the fragment environment. The point is that only those structures which meet all the requirements imposed by the environment descriptors are stored in the course of structure generation. This drastically reduces the number of structures to be checked later. On the one hand, there are problems which cannot be solved at all without descriptor control over structures during the structure generation stage (calculations take too long a time). On the other hand, there is a possibility of losing a proper structure in the case of too rigorous descriptor control. For instance, if the O=C−C=C fragment defined in the library as acyclic (whereas the proper structure contains the C=C bond in the ring) was selected at the SGA stage, all structures containing this fragment in the ring (including the proper structure) are rejected in the course of structure generation. This leads to an incorrect solution. Ideally, it is desired to have an ES characterised by high selectivity at SGA stage and, at the same time, by the capability to generate unambiguous and correct solutions. It is hardly probable to meet both requirements, therefore a compromise should be found.

Since the basic requirement for the solution is that it should be correct, it is possible to put up with its ambiguity. Therefore, the optimum strategy of structure elucidation can be formulated as follows: the problem should be solved in the interactive mode, which is a sequence of iterations followed by increasing the hardness of imposed constraints. For instance, if dimensionality of the problem is not very large, the library of principal functional groups (LPF, see Section V.1) should be used to select the fragments in the first stage. This provides the possibility of using softest structural constraints to generate a maximum number of fragment environments. Recognition of a large molecule requires additional use of the SGA and AROM libraries.

To control the hardness of imposed constraints at the stages of fragment selection and structure generation, the chemist can switch on (or off) the following options: the use of any type of spectra from the available experimental data; the use of signal multiplicities and SSCC in NMR spectra; the use of band intensities and halfwidths in IR spectra; the environment descriptor control over structures in the course of structure generation; the environment descriptor control over structures in the course of structure filtration; setting of various tolerances for the predicted number of signals with different multiplicity in the ^{13}C NMR spectrum; the use of structural constraints at the stages of structure generation and filtration (the limiting size of the ring, the maximum bond multiplicity, satisfying Bredt's rule, GOODLIST, BADLIST, list of unlikely fragments, *etc.*).

If no solution was obtained during the first program run which, as a rule, is carried out automatically, rational use of the above-mentioned switchable parameters in the course of iterative (trial-and-error) procedure usually makes it possible to establish the reasons for contradictions, thus facilitating the search for a correct solution. The search for contradictions is particularly efficient when using the controlled filtration mode (see Section V.1). Storing all intermediate results in the active database allows the user to return to the solution of the problem at any step.

Fast operation of programs for prediction of ^{13}C NMR spectra (0.5–1 s per spectrum) makes possible effective employment of the strategy outlined above. Thus, to attain a correct solution, the initial constraints may sometimes be slacked to such a degree that the answer file will contain thousands of structures. Prediction of the chemical shifts and the maximum likelihood ranking of the structures are carried out automatically; therefore CPU time taken can be considered justified and admissible.

The choice of appropriate strategy of solving a certain problem depends on optimum combination of two basic modes of operation of the system, which can be called the 'Archaeologist' mode and the 'Sculptor' mode. The 'Archaeologist' mode is used to perform SGA and structure generation (to 'glue' fragments and atoms). Then the 'Sculptor' mode should be switched on to carry out structure filtration and reject wrong structures (in this mode, the strategy is to cut off all superfluous structures).

Analysis of a large number of solutions made it possible to divide all problems into the following categories:

(1) problems solved automatically using SSG from molecular formulae followed by application of the 'Sculptor' mode strategy;

(2) problems solved automatically using strategies of the 'Archaeologist' and 'Sculptor' modes;

(3) problems solved using elements of the interactive mode (a chemist edits the 'reasoning' of the program using diversified information easily obtained from experimental data); and

(4) problems which can be solved only in the interactive mode (manual creation of fragment sets, imposing various constraints due to the origin of a specimen, *etc*.).

Rating a problem in a particular category depends on different factors (*e.g*., the size of the molecule, the presence of large fragments with characteristic spectral features, symmetry, *etc*.) and, first of all, on the number of skeletal atoms in the empirical formula. In many instances, SGA can be performed automatically; however, the appearance of at least one contradiction between the axioms of the SGA theory and experimental data makes the interference of an expert virtually inevitable. It should be kept in mind that even coincidence of predicted spectra with experimental ones does not guarantee against incorrect choice of the structure. Additional experiments can appear to be necessary, in particular, an X-ray study or structure confirmation by INADEQUATE 2D NMR technique. Despite carrying out additional labour-consuming procedures, the expenses are incommensurate with the cost of independent synthesis, isolation and purification of the product.

Thus, AI programs must be considered as powerful tools of extending the intellectual potentialities of the investigator rather than those of a method designed to exclude a chemist from the structure elucidation process. This idea was clearly expressed by the authors of monograph:[170] 'Computer-based automated or interactive versions of similar approaches have also been devised for structural elucidation of complex natural products, such as the SESAMI (Systematic Elucidation of Structures by using Artificial Machine Intelligence), but there is no substitute for the hard work, experience and intuition of the chemist.' We cannot but agree with this.

VII. Conclusion

The first studies reported 30 years ago laid the foundations of an independent line of investigation, the development of ES for elucidation of the molecular structure of organic compounds by spectral methods. It is closely related to many branches of chemistry (organic, physical, analytical, quantum chemistry, stereochemistry, molecular spectroscopy, molecular mechanics), mathematics (theory of graphs, combinatorial analysis, mathematical logic, theory of fuzzy sets, mathematical statistics), cybernetics (AI, programming, information theory, pattern recognition, ANN theory, computer graphics, mathematical linguistics), psychology (the man–computer interface) and philosophy of science (the role of the complementarity principle in studying the structure of matter).[171] Therefore rapid progress in the development of ES that absorbs new achievements of interdisciplinary sciences should be expected.

Three major stages of the process of structure recognition, *viz*., the SGA, structure generation and prediction of the spectra of candidate structures, have been transformed into independently developing investigation lines. It has been possible to computerise the structure generation and prediction of NMR spectra to a greater extent than SGA. Several methods were developed to reveal molecular fragments; however, this stage remains the most vulnerable and dependent on the *a priori* information. Recently, intensive studies on the application of ANN to revealing the fragments have begun.

Up to now, all attempts to develop a fast and reliable method for the prediction of IR spectra have failed. Solving this problem requires investigation of the potentialities of ANN (encouraging results have already been obtained) and development of new approaches. The known methods of calculation of IR spectra must be essentially improved to accelerate calculations. It has been shown [172–174] that calculations of IR spectra can also be used for quantitative spectral analysis of organic compounds without any calibration. Therefore incorporation of databases (*e.g*., such as it was made by Gribov and Zinovyev [138]) into ES will allow one to create systems for both qualitative and quantitative spectral analysis of organic substances.

Taking into account the accumulated experience in solving many structural problems, the general strategy of elucidation of the molecular structure using ES was established. It consists of performing a number of successive iterations followed by a gradual increase in the hardness of constraints imposed. The use of this strategy makes it possible to obtain correct solutions.

Currently, most of ES are capable of recognising small and medium size (up to 20–25 skeletal atoms) molecules; however, this does not mean that ES have limited area of application. Millions of organic structures can correspond to a relatively simple molecular formula (*e.g*., about 48.5 mln of structures correspond to the empirical formula $C_9H_{10}O_3$); since 'only' $\sim$ 14 mln of compounds are known currently, it is easy to conclude that the area of application of ES is very wide.

Since at present the structures generated by the use of the most sophisticated structure generators contain no more than 10–12 molecular fragments, the ES capable of revealing large fragments are rather promising for recognition of large molecules. Rapid progress of ES in which 2D NMR spectroscopy data are used should also be expected.

References

1. M E Elyashberg, L A Gribov *Zh. Prikl. Spektrosk*. **8** 296 (1968)
2. M E Elyashberg, L A Moskovkina *Zh. Prikl. Spektrosk*. **8** 996 (1968)
3. J Lederberg, G L Sutherland, B G Buchanan, E A Feigenbaum, A V Robertson, A M Duffield, C Djerassi *J. Am. Chem. Soc*. **91** 2973 (1968)
4. S I Sasaki, H Abe, T Ouki, M Sakamoto, S Ochiai *Anal. Chem*. **40** 2220 (1968)
5. D B Nelson, M E Munk, K B Gasli, D L Horald *J. Org. Chem*. **34** 3800 (1969)
6. V M Tatevskii *Klassicheskaya Teoriya Stroeniya Molekul i Kvantovaya Mekhanika* (Classical Theory of Structure of Molecules and Quantum Mechanics) (Moscow: Khimiya, 1973)
7. M E Elyashberg, L A Gribov, V V Serov *Molekulyarnyi Spektral'nyi Analiz i EVM* (Molecular Spectral Analysis and Computers) (Moscow: Nauka, 1980)
8. R K Lindsay, B G Buchanan, E A Feigenbaum, J Lederberg *Applications of Artificial Intelligence for Organic Chemistry. The DENDRAL Project* (New York: McGraw-Hill, 1980)
9. N A B Gray *Computer-Assisted Structure Elucidation* (New York: Wiley, 1986)
10. T H Pierce, B A Hohne *Artificial Intelligence Applications in Chemistry* (Washington, DC: American Chemical Society, 1986)

11. J Klassens, G Kateman *Fresenius Z. Anal. Chem.* **326** 203 (1987)
12. N A B Gray *Anal. Chim. Acta* **210** 9 (1988)
13. G W Small *Anal. Chem.* **59** 535A (1987)
14. M E Elyashberg *Matematicheskie Metody i EVM v Analiticheskoi Khimii* (Mathematical Methods and Computers in Analytical Chemistry) (Moscow: Nauka, 1989) p. 132
15. M E Elyashberg, V V Serov, L A Gribov *Talanta* **34** 21 (1987)
16. L A Gribov, M E Elyashberg, V V Serov *J. Mol. Struct.* **50** 371 (1978)
17. A M Denisov *Vvedenie v Teoriyu Obratnykh Zadach* (Introduction to Theory of Inverse Problems) (Moscow: Moscow State University, 1994)
18. A N Tikhonov, V Ya Arsenin *Metody Resheniya Nekorrektnykh Zadach* (Methods of Solution of Incorrect Problems) (Moscow: Nauka, 1979)
19. V V Serov, L A Gribov, M E Elyashberg *J. Mol. Struct.* **129** 183 (1985)
20. A Kaufman *Theory of Fuzzy Set* (New York: Academic Press, 1975)
21. H J Luinge *Vibrat. Spectrosc.* **1** 3 (1990)
22. M E Elyashberg *Zh. Analit. Khim.* **47** 966 (1992) [a]
23. W A Warr *Anal. Chem.* **65** 1087A (1993)
24. J Zupan, J Gasteiger *Neural Networks for Chemists* (Weinheim: VCH, 1993)
25. E W Robb, M S Madison, M E Munk *Mikrochim. Acta.* **II** 505 (1991)
26. C Affolter, J T Clerc *Chemometr. Intell. Lab. Syst.* **21** 121 (1993)
27. D Ricard, C Cachet, D Cabrol-Bass, T P Forrest *J. Chem. Inf. Comput. Sci.* **33** 202 (1993)
28. M A Sharaf, D L Illman, B R Kowalski *Chemometrics* (New York: Wiley, 1986)
29. M E Elyashberg, E R Martirosian, Yu Z Karasev, H Thiele, H Somberg *Anal. Chim. Acta* **337** 265 (1997)
30. M Will, W Fachinger, J R Richert *J. Chem. Inf. Comput. Sci.* **36** 221 (1996)
31. *ACD/Structure Elucidator* (Advanced Chemistry Development Inc., www.acdlabs.com)
32. C Peng, S Yuan, C Zheng, Y Hui *J. Chem. Inf. Comput. Sci.* **34** 805 (1994)
33. L A Gribov, M E Elyashberg *J. Mol. Struct.* **5** 179 (1970)
34. V V Serov, M E Elyashberg, L A Gribov *Dokl. Akad. Nauk SSSR* **232** 592 (1977) [b]
35. M E Elyashberg, L A Gribov, V N Koldashov, I V Pletnev *Dokl. Akad. Nauk SSSR* **268** 112 (1983) [b]
36. M E Elyashberg, V V Serov, E R Martirosian, L A Zlatina, Yu Z Karasev, V N Koldashov, Yu Yu Yampolskiy *J. Mol. Struct.* **230** 191 (1991)
37. K Funatsu, Y Susuta, S Sasaki *Anal. Chim. Acta* **201** 55 (1989)
38. G N Andreev, O K Argirov, P N Penchev *Anal. Chim. Acta* **284** 131 (1993)
39. G N Andreev, O K Argirov *J. Mol. Struct.* **347** 439 (1995)
40. V V Serov, M E Elyashberg *Zh. Strukt. Khim.* **27** (2) 32 (1986) [c]
41. L J Bellamy *The Infrared Spectra of Complex Molecules* (Chichester: Wiley, 1975)
42. R R Yager (Ed.) *Fuzzy Set and Possibility Theory* (New York: Pergamon Press, 1982)
43. C L Chang, R C T Lee *Symbolic Logic and Mechanical Theorem Proving* (New York: Academic Press, 1973)
44. V V Serov, M E Elyashberg, V M Petrov *Matematicheskie Metody i EVM v Analiticheskoi Khimii* (Mathematical Methods and Computers in Analytical Chemistry) (Moscow: Nauka, 1989) p. 150
45. H J Luinge, G J Kleywegt, H A van't Klooster, J H van der Maas *J. Chem. Inf. Comput. Sci.* **27** 95 (1987)
46. H J Luinge, J N van der Maas *Anal. Chim. Acta* **223** 135 (1989)
47. H J Luinge *Trends Anal. Chem.* **9** 66 (1990)
48. H Huixiao, X Xinquan *J. Chem. Inf. Comput. Sci.* **30** 203 (1990)
49. C Cleva, C Cachet, D Cabrol-Bass, T P Forrest *Anal. Chim. Acta* **348** 255 (1997)
50. C Klawun, C L Wilkins *J. Chem. Inf. Comput. Sci.* **36** 69 (1996)
51. M E Munk, M S Madison, E W Robb *J. Chem. Inf. Comput. Sci.* **36** 231 (1996)
52. M E Munk, M S Madison, E W Robb *Mikrochim. Acta* **II** 505 (1991)
53. N Sbirrazzuoli, C Cachet, D Cabrol-Bass, T P Forrest *Neural Comput. Applic.* **1** 229 (1993)
54. J R M Smith, P Schoenmakers, A Stehmann, F Sijstermans, G Kateman *Chemometr. Intell. Lab. Syst.* **18** 27 (1993)
55. C Klawun, C L J Wilkins *Chem. Inf. Comput. Sci.* **34** 984 (1994)
56. E W Robb, M E Munk *Mikrochim. Acta* I 131 (1990)
57. R J Fessenden, L Gyorgi *J. Chem. Soc., Perkin Trans. 2* 1755 (1991)
58. O C van Est, P Schoenmakers, J R M Smits, W P M Nijssen *Vib. Spectrosc.* **4** 263 (1993)
59. M Meyer, T Weigelt *Anal. Chim. Acta* **265** 183 (1992)
60. M Meyer, H Hobert *Anal. Chim. Acta* **282** 407 (1993)
61. U M Weigel, R J Herges *J. Chem. Inf. Comput. Sci.* **32** 723 (1992)
62. B J Hare, J H Prestegard *J. Biomol. NMR* **4** 35 (1994)
63. C Klawun, C L Wilkins *J. Chem. Inf. Comput. Sci.* **36** 249 (1996)
64. I I Strokov, I V Gritsenko, K S Lebedev *Izv. Sib. Otd. Akad. Nauk. Ser. Khim. Nauk* (2) 78 (1987)
65. K S Lebedev, E A Otmakhova, I V Gritsenko *Izv. Sib. Otd. Akad. Nauk. Ser. Khim. Nauk* (5) 73 (1990)
66. K S Lebedev, O N Sharapova, I K Korobeinicheva, V A Kokhov *Sib. Khim. Zh.* (1) 50 (1993)
67. K S Lebedev *Zh. Analit. Khim.* **48** 851 (1993) [a]
68. K S Lebedev, I V Gritsenko *Zh. Strukt. Khim.* **34** (3) 51 (1993) [c]
69. K S Lebedev, B G Derendyaev *Khim. Interesakh Ustoichivogo Razvitiya* **3** 269 (1995)
70. K S Lebedev, D Cabrol-Bass *J. Chem. Inf. Comput. Sci.* **38** 410 (1998)
71. K Varmuza, P N Penchev, H Scsibrany *J. Chem. Inf. Comput. Sci.* **38** 420 (1998)
72. L Chen, W Robien *J. Chem. Inf. Comput. Sci.* **34** 934 (1994)
73. W Bremser *Anal. Chim. Acta* **103** 355 (1978)
74. M Carabedian, I Dagane, J E Dubois *Anal. Chem.* **60** 2186 (1988)
75. V V Raznikov, V L Tal'roze *Zh. Strukt. Khim.* **11** 357 (1970) [c]
76. Y Kudo, S Sasaki *J. Chem. Inf. Comput. Sci.* **16** 43 (1976)
77. V V Serov, M E Elyashberg, L A Gribov *Dokl. Akad. Nauk SSSR* **224** 109 (1975) [b]
78. C A Shelley, M E Munk *Anal. Chim. Acta* **133** 507 (1981)
79. B D Christie, M E Munk *J. Chem. Inf. Comput. Sci.* **28** 87 (1988)
80. A Kerber, R Laue, D Moser *Anal. Chim. Acta* **235** 221 (1990)
81. C Benecke, R Grund, A Kerber, R Laue, T Wieland *J. Chem. Inf. Comput. Sci.* **72** 403 (1995)
82. C Benecke, R Grund, R Hohberger, A Kerber, R Laue, T Wieland *Anal. Chim. Acta* **314** 141 (1995)
83. T Wieland, A Kerber, R Laue *J. Chem. Inf. Comput. Sci.* **36** 413 (1996)
84. S G Molodtsov *Commun. Math. Chem. (MATCH)* **30** 213 (1994)
85. S G Molodtsov *Commun. Math. Chem. (MATCH)* **30** 203 (1994)
86. S G Molodtsov *Commun. Math. Chem. (MATCH)* **37** 157 (1998)
87. R E Carhart, D H Smith, N A B Gray, J G Nourse, C Djerassi *J. Org. Chem.* **46** 1709 (1981)
88. S G Molodtsov, Candidate Thesis in Chemical Sciences Novosibirsk Institute of Organic Chemistry, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 1997
89. M S Molchanova, Candidate Thesis in Chemical Sciences Institute of Organic Chemistry, Russian Academy of Sciences, Moscow, 1997
90. L M Masinter, N S Sridharan, J Lederberg, D H Smith *J. Am. Chem. Soc.* **96** 7702 (1974)
91. M S Molchanova, V V Shcherbukhin, N S Zefirov *J. Chem. Inf. Comput. Sci.* **36** 888 (1996)
92. Ch-L Hu, L Xu *Anal. Chim. Acta* **298** 78 (1994)
93. J-M Nuzillard, W Naanaa, S Pimont *J. Chem. Inf. Comput. Sci.* **35** 1068 (1995)
94. R Barone, F Barberis, M Chanou *Commun. Math. Chem. (MATCH)* **32** 19 (1995)
95. I P Bangov *J. Chem. Inf. Comput. Sci.* **34** 318 (1994)
96. V Kvasnicka, J Pospichal *J. Math. Chem.* **9** 181 (1992)
97. V Kvasnicka, J Pospichal *J. Chem. Inf. Comput. Sci.* **36** 516 (1996)
98. L A Zlatina, M E Elyashberg *Zh. Strukt. Khim.* **32**(4) 92 (1991) [c]
99. L A Zlatina, M E Elyashberg *Commun. Math. Chem. (MATCH)* **27** 191 (1992)
100. E V Gordeeva, A R Katrizky, V V Shcherbukhin, N S Zefirov *J. Chem. Inf. Comput. Sci.* **33** 102 (1993)
101. J Gasteiger, C Rudolph, J Sadowski *Tetrahedron. Comput. Methodol.* **3** 537 (1990)
102. J Sadowski, J Gasteiger *Chem. Rev.* **93** 2567 (1993)
103. J Sadowski, J Gasteiger, G Klebe *J. Chem. Inf. Comput. Sci.* **34** 1000 (1994)

104. C A Shelley, M E Munk *Anal. Chem.* **50** 1522 (1978)
105. J T Clerc, H A Sommerauer *Anal. Chim. Acta* **95** 33 (1977)
106. L Chen, W Pobien *Anal. Chem.* **65** 12282 (1993)
107. K L Jensen, A S Barber, G W Small *Anal. Chem.* **63** 1082 (1991)
108. D L Clouser, P C Jurs *Anal. Chim. Acta* **295** 221 (1994)
109. P C Jurs, J W Ball, L S Anker, T L Friedman *J. Chem. Inf. Comput. Sci.* **32** 272 (1992)
110. A Furst, E Pretsch *Anal. Chim. Acta* **229** 17 (1990)
111. W Bremser *Magn. Reson. Chem.* **23** 271 (1985)
112. L Chen, W Robien *Chemometr. Intell. Lab. Syst.* **19** 217 (1993)
113. C W Crandell, N A B Gray, D H Smith *J. Chem. Inf. Comput. Sci.* **22** 48 (1982)
114. H N Cheng, L J Kasehagen *Anal. Chim. Acta* **285** 223 (1994)
115. C W von der Lieth, J Seil, I Kohler, H J Opferkuch *Magn. Reson. Chem.* **23** 1048 (1985)
116. A Furst, E Pretsch, W Robien *Anal. Chim. Acta* **233** 213 (1990)
117. R C Schweitzer, G W Small *J. Chem. Inf. Comput. Sci.* **37** 249 (1997)
118. *ACD/CNMR Predictor* (Advanced Chemistry Development Inc., www.acdlabs.com)
119. V Kvasnicka *J. Math. Chem.* **6** 63 (1991)
120. J P Doucet, A Panaye, E Feuilleaubois, P Ladd *J. Chem. Inf. Comput. Sci.* **33** 320 (1993)
121. A Panaye, J P Doucet, B Fan, E Feuilleaubois, S R El Azzouzi *Chemometr. Intell. Lab. Syst.* **24** 129 (1994)
122. V Kvasnicka, S Sklenak, J Pospichal *J. Chem. Inf. Comput. Sci.* **32** 742 (1992)
123. L S Anker, P C Jurs *Anal. Chem.* **64** 1157 (1992)
124. Y Miyashita, H Yoshida, O Yaegashi, T Kimura, H Nishiyama, S Sasaki *J. Mol. Struct. (Theochem)* **311** 241 (1994)
125. B E Mitchell, P C Jurs *J. Chem. Inf. Comput. Sci.* **36** 58 (1996)
126. D L Clouser, P C Jurs *J. Chem. Inf. Comput. Sci.* **36** 168 (1996)
127. O Ivanciuc, J-P Rabine, D Cabrol-Bass, A Panaye, J P Doucet *J. Chem. Inf. Comput. Sci.* **36** 644 (1996)
128. O Ivanciuc, J-P Rabine, D Cabrol-Bass, A Panaye, J P Doucet *J. Chem. Inf. Comput. Sci.* **37** 587 (1996)
129. D M Grant, E G Paul *J. Am. Chem. Soc.* **86** 2984 (1964)
130. E Breitmaier, W Voelter *Carbon-13 NMR Spectroscopy* (Weinheim: VCH, 1987)
131. E Pretsch, T Clerc, J Seibl, W Simon *Tables of Spectral Data for Structure Determination of Organic Compounds* (Berlin: Springer, 1989)
132. N A B Gray, J G Nourse, C W Crandell, D H Smith, C Djerassi *Org. Magn. Res.* **15** 375 (1981)
133. K R Beebe, B R Kowalski *Anal. Chem.* **59** 1007A (1987)
134. R B Shaller, E Pretch *Anal. Chim. Acta* **133** 507 (1981)
135. R B Shaller, M E Munk, E Pretch *J. Chem. Inf. Comput. Sci.* **36** 239(1996)
136. C A Shelley, M E Munk *Anal. Chim. Acta* **296** 295 (1994)
137. *ACD/HNMR* (Advanced Chemistry Development Inc., www.acdlabs.com)
138. L A Gribov, K A Zinovyev *J. Mol. Struct.* **268** 191 (1992)
139. L A Gribov, W J Orville-Thomas *Theory and Methods of Calculation of Molecular Spectra* (Chichester: Wiley, 1988)
140. M E Elyashberg, Yu Z Karasev, V A Dement'ev, L A Gribov *Interpretirovannye Kolebatel'nye Spektry Uglevodorodov — Proizvodnykh Tsiklogeksana i Tsiklopentana* (Interpreted Vibrational Spectra of Hydrocarbons — Derivatives of Cyclohexane and Cyclopentane) (Moscow: Nauka, 1988)
141. M E Elyashberg, E R Martirosian, Yu Z Karasev, H Thiele, H Somberg *Anal. Chim. Acta* **348** 443 (1997)
142. J E Dubois, G Mathieu, P Peguet, A Panaye, J P Doucet *J. Chem. Inf. Comput. Sci.* **30** 290 (1990)
143. Ch Affolter, J T Clerc *Chemometr. Intell. Lab. Syst.* **21** 151 (1993)
144. U M Weigel, R J Herges *Anal. Chim. Acta* **331** 63 (1996)
145. J H Schuur, P Selzer, J Gasteiger *J. Chem. Inf. Comput. Sci.* **36** 334 (1996)
146. J Gasteiger, J Sadowski, J Schuur, P Selzer, L Steinhauer, V Steinhauer *J. Chem. Inf. Comput. Sci.* **36** 1030 (1996)
147. J Gasteiger, J Schuur, P Selzer, L Steinhauer, V Steinhauer *Fresenius J. Anal. Chem.* **359** 50 (1997)
148. K Baumann, J T Clerc *Anal. Chim. Acta* **348** 327 (1997)
149. G W Adamson, M F Lynch, W G Town *J. Chem. Soc., C* 3702 (1970)
150. J T Clerc, A Tercovics *Anal. Chim. Acta* **235** 93 (1990)
151. P Willett, J M Barnard, G M Downs *J. Chem. Inf. Comput. Sci.* **38** 983 (1998)
152. N A B Gray *Chemometr. Intell. Lab. Syst.* **5** 11 (1988)
153. B G Buchanan, E A Feigenbaum, J Lederberg *Chemometr. Intell. Lab. Syst.* **5** 33 (1988)
154. N A B Gray *Chemometr. Intell. Lab. Syst.* **5** 37 (1988)
155. M E Munk *J. Chem. Inf. Comput. Sci.* **38** 997 (1998)
156. M E Elyashberg, Yu Z Karasev, E R Martirosian *Chimia* **52** 324 (1998)
157. W Bremser *Angew. Chem., Int. Ed. Engl.* **27** 247 (1988)
158. K Funatsu, Y Susuta, S Sasaki *J. Chem. Inf. Comput. Sci.* **29** 6 (1989)
159. K Funatsu, S Sasaki *J. Chem. Inf. Comput. Sci.* **36** 190 (1996)
160. M E Munk, B D Christie *Anal. Chim. Acta* **216** 57 (1989)
161. M E Munk, V K Velu, M S Madison, E W Robb, M Christie, M Razinger, in *Recent Advances in Chemical Information II* (Ed. H Collier) (Cambridge: Royal Society of Chemistry, 1993) p. 247
162. M Christie, M E Munk *J. Am. Chem. Soc.* **113** 3750 (1991)
163. C Peng, S Yuan, C Zheng, Y Hui, H Wu, K Ma, X Han *J. Chem. Inf. Comput. Sci.* **34** 814 (1994)
164. C Peng, S Yuan, C Zheng, Z Shi, H Wu *J. Chem. Inf. Comput. Sci.* **35** 539 (1995)
165. P L Fuchs, C A Bunnell *Carbon-13 NMR Based Organic Spectral Problems* (New York: Wiley, 1979)
166. E Pretsch, J Seibl, A Manz, J Simon *Aufgabensammlung zur Strukturaufklarung Organischer Verbindungen mit Spektroskopischen Methoden* (Berlin: Springer, 1985)
167. R B Bates, W A Beavers *Carbon-13 NMR Spectral Problems* (Clifton: Humana Press, 1981)
168. J T Clerc, E Pretsch, J Seibl *Structural Analysis of Organic Compounds by Combined Application of Spectroscopic Methods* (Budapest: Akademiai Kiadó, 1981)
169. E Pretsch, A F Furst, M Badertscher, R B Shaller, M E Munk *J. Chem. Inf. Comput. Sci.* **32** 291 (1992)
170. N F Atta-ur-Rahman, M I Choudhary *Solving Problems with NMR Spectroscopy* (New York: Academic Press, 1996) p. 391
171. L A Gribov *J. Mol. Struct.* **300** 415 (1993)
172. L A Gribov, M E Elyashberg, Yu Z Karasev *Anal. Chim. Acta* **316** 217 (1995)
173. M E Elyashberg, L A Gribov, Yu Z Karasev, E R Martirosian *Anal. Chim. Acta* **353** 105 (1997)
174. L A Gribov, D I Sidelov, I V Maslov *Zh. Analit. Khim.* **53** 706 (1998)

[a] — *J. Anal. Chem. (Engl. Transl.)*
[b] — *Dokl. Chem. Technol., Dokl. Chem. (Engl. Transl.)*
[c] — *J. Struct. Chem. (Engl. Transl.)*