

Chapter 2

Strategies of Structure Elucidation

Abstract Different strategies regarding the application of the system depending on the specific features of the problem under analysis are discussed in this chapter. The input of 1D and 2D NMR spectroscopy data (automated and manual) to the program and the creation of electronic data tables are outlined. Special attention is focused on developing a Molecular Connectivity Diagram (MCD) as well as its editing and checking for consistency. An MCD represents visually *all initial information* employed by the system for structure generation. The application of the *Common* and *Fragment* modes of *Strict Structure Generation* depends on the particular peculiarities associated with a problem (molecule size, deficit of hydrogen atoms, etc.) and the methodology of the most probable structure selection from the output file is described. Special consideration is given to the problem of resolving logical contradictions in 2D NMR data arising from the presence of “nonstandard” *correlations* (those for which ${}^nJ_{\text{HH, CH}, n > 3}$). To this aim, a *Fuzzy Structure Generation* (FSG) algorithm is implemented into Structure Elucidator which allows for identification of the structure of the compound under analysis in the presence of an *unknown number* of nonstandard correlations of *unknown lengths*. Different modes of FSG are expounded and strategies for its application are discussed, instructing the student when and how each mode can be effectively employed.

2.1 Data Input, Processing, and Forming of a Molecular Connectivity Diagram

We assume that the reader is skilled and experienced enough to manually process raw 1D and 2D NMR data using traditional approaches (peak picking, structural interpretation of spectral features, etc.). These abilities are also necessary in all stages of CASE problem solving, especially during the preparation of spectral data for input into a computer. This stage is very important and it can be considered as a first step in forming the primary “axioms” and hypotheses (Sect. 1.2). It should be strongly emphasized that the application of the CASE system does not release the

chemist from having the necessary knowledge and experience in NMR spectroscopy. At the same time the program takes responsibility for the automatic creation of the majority of axioms and hypotheses necessary for structure elucidation, leaving their approval to the chemist. Once the initial set of “axioms” and hypotheses is adopted by the chemist further structure inference is performed by the system automatically. We also assume that the chemist has knowledge and experience in mass spectrum interpretation. Skills in the analysis of a peak cluster observed around the molecular ion and in deducing the molecular formula from an accurate molecular mass are crucially important for utilization of a CASE system. Therefore, in this chapter we will briefly describe the main steps of initial NMR spectral data preprocessing and data input into the program using the facilities and interface developed for Structure Elucidator. For this goal, a typical example will be used where the 1D and 2D spectra are of good quality and the molecule under investigation is a natural product of common complexity.

In the examples presented in Part III of this textbook, we will use 1D and 2D NMR data that were already processed and saved as electronic tables in the format common for Structure Elucidator. Primary attention will be placed on the methods of overcoming uncertainty, incompleteness, and contradictions in the initial data and the different operating modes provided in Structure Elucidator for these goals will be explained.

2.1.1 Data Used for Structure Elucidation

There are various types of data that can be used to perform computer-assisted structure elucidation. In particular, as described in the previous sections, 2D NMR is the essential technique for the elucidation of complex chemical structures and, due to its inherent complexity in terms of processing and manipulation, it is this form of NMR that puts significant demands on the software. The 2D NMR structure generator requires as input a set of atoms and the connectivities between them. The generator also takes into account the associated chemical shifts of the atoms as well as a series of different structural constraints. The base set of atoms is usually obtained from the molecular formula, while the connectivities between the atoms are revealed by combining the data encoded into the 1D and 2D NMR spectra.

The following spectra are commonly used in structure elucidation:

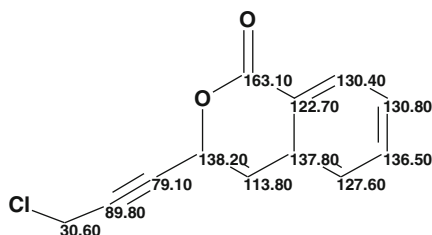
- **Mass spectra** mainly provide the molecular mass and elemental composition (in combination with NMR and IR spectra) and, as a result, access to the molecular formula for the compound under examination. Note that the determination of a molecular formula is crucial for the structure elucidation both manually or using an expert system. The fragmentation of the compound represented by a mass spectrum provides access to further detail regarding the molecular composition in terms of key molecular fragments present in the compound under study.

- **1D NMR spectra** contain information about atoms included in the structure in terms of their electronic environments, their proximity to each other both in terms of skeletal connections and through-space interactions, as well as details regarding internal barriers of rotation.
- **2D NMR spectra** can provide information about both through-bond and through-space interactions between atoms and are the most informative, especially when multiple types of 2D NMR spectra are acquired and analyzed in parallel.

Data preparation is a key part of both manual and automated structure elucidation. Almost any error made during this procedure can lead to erroneous structures being derived as a result of the elucidation process. For input into Structure Elucidator, data preparation should therefore be done as carefully as possible. Data preparation consists of two main steps—the determination of an atom list and the determination of the connectivities between atoms. Algorithms for structure generation can automatically correct inconsistency and some errors in connectivities between atoms but it is almost impossible to rectify mistakes in the list of atoms. A dialog window **Spectrum Parameters** is used to specify the main spectrometer parameters which are set for corresponding spectrum registration.

The chemist has the possibility to postulate values of chemical shift **User Defined Tolerance** (ppm) for the F1 and F2 axes. These parameters significantly influence the problem complexity (size of the output file, time of structure generation, etc.).

Most procedures described in this chapter will be illustrated using the example of a small molecule, gymnopalynes [1], which contains carbon atoms of diverse properties and several heteroatoms. Gymnopalynes has the molecular formula $C_{12}H_7O_2Cl$. Its structure **2.1** with the assigned ^{13}C chemical shifts is shown below.

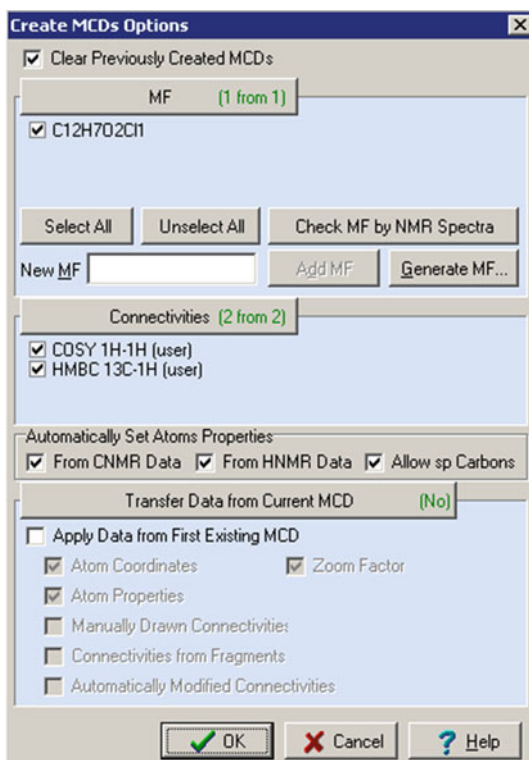


2.1.2 Molecular Formula

The structure generation algorithm requires a list of atoms as initial data input. The molecular formula (MF) is a compact representation of this list and is an absolute requirement in order to perform structure elucidation. Usually a monoisotopic

(accurate) mass is used to determine the molecular formula and is determined by analyzing the mass spectrum for the mass (m/z) of the molecular ion. It should be noted that a molecular ion peak is not always present in the mass spectrum. This is more common in electron impact ionization spectra but spectra obtained using other (more “mild”) ionization methods, for example, electrospray, usually do contain the molecular ion peak. Mass spectra obtained using the positive ion electrospray ionization (ESI) method contain the adduct peak of the protonated molecular ion. In addition to the protonated ion, usually called $[M+H]^+$, other adducts include ions such as Na^+ , K^+ or NH_4^+ , denoted as $[M+Na]^+$, $[M+K]^+$, and $[M+NH_4]^+$. This information should be taken into account when generating molecular formulae, i.e., the monoisotopic mass should be corrected as appropriate. Structure Elucidator is supplied with the Molecular Formula Generator which allows for the generation of all molecular formulae corresponding to a given molecular mass and postulated mass tolerance tol_m . To get to the **Molecular Formula Generator** it is necessary first to activate the command **Structure Elucidation/Create Molecular Connectivity Diagram**, as a result of which the dialog window **Create MCDs Options** will appear (Fig. 2.1). To open the dialog window **Molecular Formula Generator** (Fig. 2.2) it is necessary to click on the key **Generate MF**.

Fig. 2.1 The dialog window **Create MCDs Options**



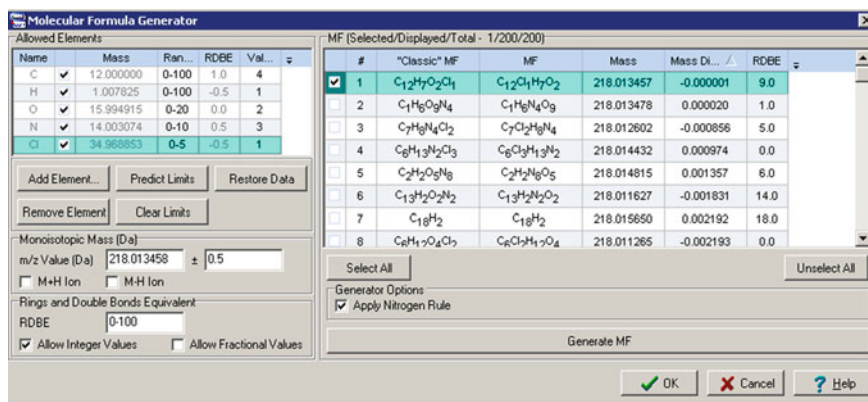


Fig. 2.2 The dialog window of the **Molecular Formula Generator**

The top of the molecular formulae list corresponding to the molecular ion of gymnopalpyne $m/z = 218.013458$ at $tol_m = 0.5$ Da are shown in the right part of the window. In the left upper corner, the chemical elements assumed to be present in a molecule are shown along with the limits of the numbers of corresponding atoms allowed by user. The limits are postulated by the user on the basis of data extracted from 1D NMR and IR spectra, as well as by taking into account the pattern of the molecular ion cluster. In particular the number of signals in the ^{13}C NMR spectrum, the values of the integrals in the ^1H NMR spectrum and the characteristic IR absorption bands observed in the region of $3,700\text{--}1,300\text{ cm}^{-1}$ serve for this goal. If no constraints regarding the number of chemical elements can be imposed, the default settings are as follows: C (0–100), H (0–100), O (0–20), N (0–10). In the left lower part of the window the monoisotopic mass and a tolerance tol_m are set. The possible limits for the RDBe value (Rings and Double Bonds Equivalent) can be input in the relevant field (0–100 as default). If chemical ionization is used to acquire the mass spectrum then the check boxes **M+H Ion** or **M–H Ion** are selected. To use the “nitrogen rule” during molecular formulae generation, the check box **Apply Nitrogen Rule** should be selected. The generation of the molecular formulae is started by clicking on the **Generate MF** key. Generated molecular formulae are displayed in the right part of the window.

The number of molecular formulae corresponding to a given monoisotopic mass can vary depending on the accuracy of the mass determination and the possible elemental composition. For example, 200 molecular formulae correspond to the monoisotopic mass of gymnopalpyne (218.013458 Da) if the accuracy of the mass determination is within 0.5 Da and only the elements C, H, N, O, and Cl are allowed. As shown in Fig. 1.17 the molecular formula of gymnopalpyne is on the top of the list (#1) because generated molecular formulae are automatically ranked in increasing order of differences. If the number of carbon atoms is restricted to 12, the corresponding number of peaks in the 1D NMR carbon spectrum, then the number of molecular formulae is reduced to 9. A similar decrease in the number of potential

formulae occurs when the accuracy of mass determination increases. When the limits of carbon atom numbers are set as the default and the accuracy is 0.05 Da then the number of molecular formulae is 107. However, when measured to an accuracy of 0.005 Da the number of formulae decreases to 8 and two molecular formulae (a true one and unrealistic formula $C_1H_6N_9O_4$) can be found when the accuracy is 0.0005 Da. Only one and correct molecular formula was generated when the tolerance was set to 0.000005. In practice both methods, the restriction of elemental composition using other data and increasing mass accuracy, are used simultaneously to identify a single molecular formula in most cases. In some relatively rare cases, when unambiguous determination of a molecular formula is impossible, the structure elucidation process can be run several times using different molecular formulae.

2.1.3 Forming the Molecular Connectivity Diagram

To provide a complete and clear pattern of the properties of the skeletal atoms and the connectivities between them the program places skeletal atoms together with hydrogen atoms attached to skeletal atoms (CH_3 , CH_2 , CH , and C groups, as well as OH and NH if identified by the user from 1H NMR and 2D NMR spectra) in a display. This pattern is referred to as the Molecular Connectivity Diagram (MCD). As mentioned above, to create the MCD it is necessary to activate the command **Structure Elucidation/Create Molecular Connectivity Diagram**. The dialog window **Create MCDs Options** (Fig. 2.1) provides for execution of the following functions:

- Use one MF or a set of possible MFs for creating the MCDs. If n formulae are selected, then n MCDs will be created.
- Input a new molecular formula directly in the field **New MF**.
- Select the types of 2D NMR data input into the program which will be used for MCD creation.
- Use the Atom Property Correlation Table (APCT) for atom property setting from the ^{13}C and 1H NMR data.
- Transfer data from the first existing MCD to a new MCD which is created by the user. Transferable properties are set using a selection of corresponding check boxes.

The MCD created from the 1H , ^{13}C , HMBC, and COSY spectra of gymnopalynes is shown in Fig. 2.3. A copy of the MCD is created simultaneously in the dialog window **Auto MCD**. All changes made in the Tables of Data are automatically transferred into the **Auto MCD** window. This window allows viewing of the initial MCD as the User MCD is edited by the user.

HMBC and COSY connectivities are shown in the MCD as “fuzzy” subgraphs (fragments) connecting carbon atoms and/or carbon and nitrogen atoms by arrows or lines when the corresponding 1H – 1H COSY and ^{15}N HMBC data are available.

sp^2 , sp , *not sp*, and “not defined”. To ease the visual recognition of the type of hybridization of a given atom each type is marked in its own specific color: sp^3 —blue, sp^2 —violet, undefined (sp^3 or sp^2)—light blue, sp —green (see Fig. 2.3). It is essential to note that both ^{13}C and ^1H NMR chemical shifts are taken into account by the program when setting the atom parameters. These descriptors for the carbon atoms allow the system to analyze 2D NMR data and to efficiently apply constraints during the process of structure generation.

If a distinct multiplet is observed in the ^1H NMR spectrum from a structural block $(\text{C}_i)\text{H}_n$ then the total number of hydrogen atoms attached to carbons adjacent to the (C_i) carbon is set. This property is determined by the chemist after visual analysis of the ^1H spectrum pattern and after taking into account the coupling constants (if measured). The atom properties should be set and edited with great caution because an erroneous assumption (a wrong “axiom”) leads to the exclusion of the correct structure from the output file. All structural constraints presented in the MCD are used during structure generation. Figure 2.4 shows a window where all properties of a particular CH_2 group are presented as an example, while Fig. 2.5 displays the pull-down menus for setting the possibility of neighboring with a heteroatom and atom hybridization.

Fig. 2.4 A window showing an example of setting the properties for a CH_2 group

Edit Properties of Atom # 1

Number of Current Atom = 1

<< Previous (15) Next (2) >>

Assigned Shift(s)
 Atom's NMR Shift (nucleus ^{13}C , shifts from -10 to 280 ppm)

Experimental 30.6 ± 0

Calculated 49.64 ± 0.000

Atom's QM calculated NMR Shift (ppm)

QM (GIAO/DFT) ±

Attached Hydrogen's NMR Shift(s) (-2 - 20 ppm)

1st Experimental 4.55 ± 0

Calculated 3.02 ± 0.000

2nd Experimental ±

Calculated n/a ± n/a

Atom Properties

Connection with Heteroatoms at least one

Number of Hydrogens on Neighbor Atoms 0

Hybridization State sp3

Charge 0

Valency not defined

☐ Allow Non-default Valences

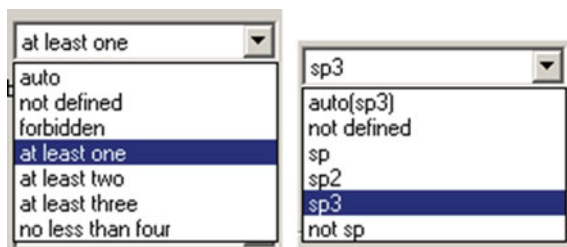


Fig. 2.5 The pull-down menus for setting the possibility of neighboring with a heteroatom (*left*) and atom hybridization (*right*)

Here a carbon atom with a chemical shift of 30.6 ppm is in a sp^3 hybridization state, and its connection with a heteroatom is obligatory. Since the signal at 4.55 ppm is a distinct singlet in the ^1H NMR spectrum the number of hydrogens attached to the carbon atoms closest to the C (30.6) carbon is set by the user to be equal to zero. The latter constraints speed up the structure generation process significantly because the generation of structures where this constraint is violated will be suppressed. The chemist can then edit these parameters using other available information. For this purpose, a set of buttons is shown on the MCD toolbar. The functions of these buttons are explained by screen tips and are intuitively clear. For example, if the molecule belongs to the CHNO class and the sp^2 hybridized carbon atom C(163.1) is marked by the user as sp^2 at least two then the system will only generate (if possible) O–C=O, N–C=O, O–C=N, and O–C \equiv N fragments on the basis of that atom. The chemist is also offered the opportunity to draw bonds of any multiplicity between the atoms to introduce suggested fragments into the process and to set some proposed functional groups (for instance C=O, O–C=O, C \equiv C, etc.). If a molecule contains heteroatoms and there are free H atoms displayed on the MCD, then O–H, N–H, NH₂, etc., groups may also be drawn in. This provides a quick and intuitive mechanism for entering structural information evident from the ^1H NMR and/or IR/Raman spectra. Lengths of connectivities can also be edited by drawing connectivities of definite lengths between selected atoms. In addition, a forbidden connectivity between a pair of atoms may be drawn on the MCD. All structural constraints presented in the MCD are used during the structure generation process. Edits of the MCD are carried out easily using the toolbar where all commands are intuitively clear and supplied with screen tips.

2.1.4 Checking the MCD for Consistency

As we will see later Structure Elucidator is capable of solving complicated problems if the spectral data are free of contradictions. The system is adjusted by default to account for the coupling constants $^{2-3}J_{\text{HH}}$ and $^{2-3}J_{\text{CH}}$ which are common for the corresponding COSY and HMBC correlations (referred to as “standard” correlations

in Sect. 1.2.2) and contradictions will appear when at least one correlation of >3 bonds results in a response in the 2D NMR data. Although the intensities of the 2D NMR peaks corresponding to “nonstandard” correlations are, in general, somewhat weaker than those corresponding to the standard correlations, the origin of both kinds of peaks is difficult to distinguish. Despite recent developments to aid in the identification of the correlation lengths [2–5] there is presently no routine NMR technique that is capable of distinguishing couplings of different lengths in a reliable fashion. Therefore, the development of theoretical methods for 2D NMR data analysis that identifies the presence of “nonstandard” correlations is of considerable importance.

The StrucEluc system is supplied with algorithms and programs [6] that are able to detect the presence of nonstandard correlations in 2D NMR data in the majority of cases. Algorithms that help to remove contradictions by lengthening certain connectivities have been delivered. A more general method to overcome the presence of contradictions in 2D NMR data, referred to as Fuzzy Structure Generation (FSG) (see Sect. 2.3), was also developed [7] and implemented into StrucEluc. In any case, the first step of structure elucidation using StrucEluc is to check the MCD for the presence or absence of contradictions in the 2D NMR data, i.e., to check data for consistency. The data checking algorithm is based on logical analysis of the full set of connectivities derived from the available 2D NMR spectra. The algorithm is sophisticated and is based on utilizing a method of logical proof by reduction *ad absurdum*. For instance, an indication of the presence of contradictions in two-dimensional data can serve a conclusion: the data could be considered consistent only in those cases where the valence of at least one carbon atom was assumed to be five or six, which is impossible. Because the algorithm is based on some heuristic statements it gives no guarantee that contradictions will be detected in any case. Experience has shown that analysis is successful in approximately 90 % of those problems where nonstandard connectivities existed in the data.

For MCD checking the command **Structure Elucidation/Check Current MCD...** is activated in the menu **Structure Elucidation** (Fig. 2.6).

As a result the dialog window **Check MCDs Options** (Fig. 2.7) is opened. Typical options which are used for the first program run are displayed in Fig. 2.7.

The options should correspond to the conditions and assumptions postulated by the user as being true during structure generation. If the check box **Automatically Resolve Contradictions** is selected the program will try to elongate all connectivities emanating from “suspicious” atoms by one bond. The relevance of the other check boxes is easily interpreted, but some explanations are necessary. The Atom Property Correlation Table (APCT) is commonly used for automatic atom property setting with “standard intervals” as shown in Fig. 2.7. The “wide intervals” can also be selected by the user or, if necessary, the APCT may be switched off by selecting the option “not used” (see Fig. 2.8, left part). For 2D NMR spectrum processing the option **“Real Spectrum”** is selected when real problems are solved (Fig. 2.8, right part).

Fig. 2.6 The **Structure Elucidation** menu



Fig. 2.7 The dialog window for **Check MCDs Options**

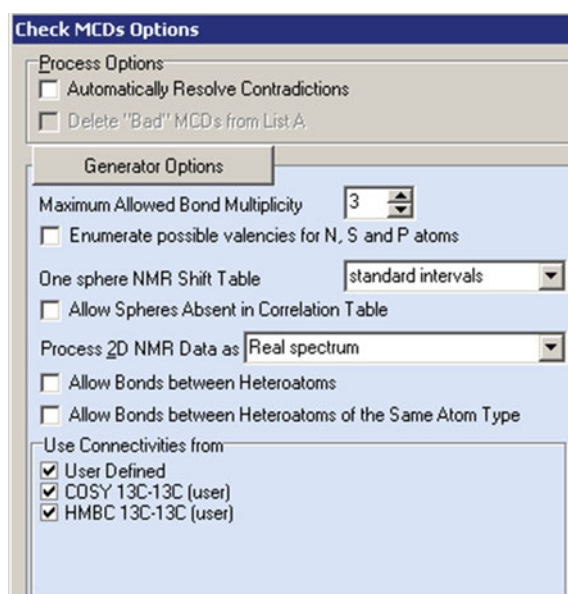




Fig. 2.8 The options of APCT (*left*) and possible selections for the field **Process 2D NMR** (*right*)

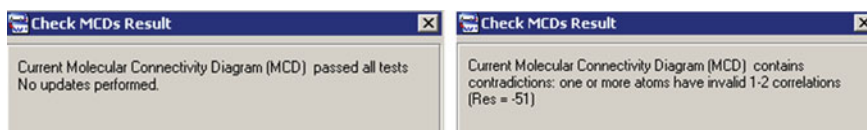


Fig. 2.9 **Check MCD Results** messages. The *left* message is delivered by the program if no contradictions were detected in 2D NMR data. If contradictions were detected the *right* message appears

When MCD checking is completed (it usually takes 1–3 s) the program displays a message containing information about the presence or absence of contradictions in the 2D NMR data. Examples of **Check MCD Results** messages are shown in Fig. 2.9.

When pressing on the key **More...** the user can see information about “suspicious” atoms and program suggestions about the minimum number of nonstandard connectivities. If no contradictions were detected then Strict Structure Generation can be performed, otherwise some different mode of the structure elucidation software is utilized (see Sect. 2.3).

A situation can be realized, however (see examples in Chap. 5), when generation of structures corresponding to all of the defined connectivities, including the non-standard ones, is *possible*. Generation is possible because the program fails to identify the nonstandard connectivities. This leads to either an *invalid* solution or to a valid solution in which chemical shift assignment for the right structure is incorrect. Therefore it can happen, even after seemingly successful checking of the data for contradictions, that nonstandard connectivities will nevertheless remain unnoticed by the algorithm. Their presence can only be identified a posteriori in indirect ways, for which the following conditions exist:

- large value(s) for the ^{13}C NMR spectral deviations calculated for the most probable structure(s);
- inconsistencies between the most probable structure and additional experimental data (for instance, NOESY, ROESY, etc.);
- disagreement between the chemical shifts and multiplicities of the experimental and calculated data of ^1H NMR spectra;
- the structures contradict IR/Raman correlations and/or interpretation of a mass spectrum.

The nonstandard connectivities that do not prevent the structure from being built are termed *implicit* nonstandard connectivities.

As the identification of implicit nonstandard connectivities cannot be guaranteed, since any of the given connectivities may be nonstandard, a method for removing nonstandard connectivities was developed which in the majority of cases allows for the identification of a valid solution even in this situation. As the method will be described in detail in Sect. 2.3, here we will explain only the main idea of the approach allowing the problem of implicit nonstandard connectivities to be circumvented.

Some connectivities are declared, as a series, as *suspicious* and are lengthened or eliminated. Each time structure generation is initiated with a renewed connectivity set. This process is termed as *Fuzzy Structure Generation* (FSG).

Let n be the total number of connectivities in the 2D NMR data and m be the number of connectivities that are suggested to be nonstandard. In this case it is necessary to consider $N = \binom{n}{m} = \frac{n!}{m!(n-m)!}$ different sets of m connectivities that will be declared as suspicious. If all N combinations of connectivities were used for structure generation, then the calculation time would increase dramatically as the number of tasks resulting in structure generation sharply increases with the rise in the m value.

Declaring members of the connectivity sets to be suspicious is used to search for atoms and pairs of atoms with nonstandard connectivities as well as for the direct determination of the presence of nonstandard connectivities. In these cases the program usually lengthens or deletes all connectivities belonging to the atoms selected during data analysis. In so doing, both nonstandard and standard connectivities are deliberately lengthened or deleted, which correspondingly leads to an increase in the number of structures generated. If only the definite connectivities (related to atoms for which the presence of nonstandard connectivities are revealed), are lengthened or deleted, then the number of generated structures will be considerably lower. The methodology and strategy of FSG will be discussed in Sect. 2.3.

2.2 Modes of the Structure Generation

In this section we will consider the main modes of the StrucEluc system. The program is capable of elucidating the chemical structure of much larger molecules, up to a mass of 1,500 amu to date, and containing more than 100 skeletal atoms. Typically, this task is accomplished from the analysis of 2D NMR spectral data. In general the system has been designed to elucidate structures containing up to 250 skeletal atoms. The capabilities of the *StrucEluc* system in terms of general utility as a tool for the structure elucidation of complex molecules, especially natural

products, has been demonstrated in many publications [8–12] (see reviews [11, 13] and the monograph [14]).

It should be emphasized that a large number of problems can be solved using only a molecular formula, heteronuclear (HSQC/HMQC, HMBC, or COLOC) and homonuclear (H–H COSY) 2D NMR correlations and without using any additional structural information. In this mode of operation, referred to as the *Common* mode, the system creates connectivities from the spectral data and generates all possible structures in accordance with the default settings for the number of intervening bonds between corresponding skeletal atoms and with atom properties including the state of hybridization and the possibility of taking neighboring heteroatoms into account.

However, it turned out that there were problems that could not be solved or proved to be very time-consuming due to a lack of information in the 2D NMR data (see, for example, Sect. 4.34). In these cases it proved necessary to introduce additional structural information, if available, to facilitate the elucidation process. In the real world, it is common for a chemist or spectroscopist faced with elucidating a structure to have prior knowledge of reaction components in a synthesis, knowledge of the class of compounds that may have been isolated, or even hypothetical structures for validation rather than full elucidation from no information.

It has been shown [14] that the utilization of molecular fragments found from the system knowledge base, or potential substructures proposed by the chemist, can be helpful to circumvent the difficulties. Such a *fragment approach* has been used in a number of first-generation expert systems based on correlation tables containing substructures and their associated characteristic intervals for specific spectral features. In contrast, StrucEluc employs a database containing substructures and their associated ^{13}C NMR subspectra. At present the StrucEluc database contains more than 290,000 chemical structures and more than two million substructures. The database continues to grow as further literature data is added. The value of including substructures directly into the elucidation process is that a fragment, considered as a macro atom, can absorb a significant number of the skeletal atoms and leads to a reduction in the complexity of the problem. This results in acceleration of the structure generation procedure, which is typically the most time-consuming stage of the structure elucidation process.

Nevertheless, in those cases when 2D NMR data is employed, the usage of molecular fragments is hampered by the fact that *all* carbon atoms existing in a fragment utilized in solving the problem *must* be supplied with chemical shifts. Moreover, the values of these chemical shifts must be as close as possible to the observed values for the atoms of the corresponding fragments in the experimental ^{13}C NMR spectrum of the unknown under study. Particularly, the approximate chemical shift values of carbon atoms can be found using ACD/C NMR Predictor. Before structure generation all approximate chemical shift values set for fragment carbon atoms should be replaced by experimental chemical shifts closest to those ascribed to fragment atoms. The reason for this requirement is obvious: the utilization of 2D NMR correlations implies the possibility to use only observed experimental chemical shifts. The accommodation of one or more fragments within

a set of connectivities derived from the 2D NMR data is a problem that requires the development of new algorithms. In this chapter we will discuss different strategies for applying the StrucEluc system. Depending on the initial data available, and the complexity of the molecule being analyzed, the system offers a wide range of methods for solving a problem.

2.2.1 The “Common” 2D NMR Mode

The *StrucEluc* system is based on a number of programs developed for elucidating a molecular structure from a combination of 2D NMR spectra. The most typical combination providing the basis for structure determination includes H–H COSY, HSQC/HMQC, and HMBC. The StrucEluc system also operates with additional 2D NMR methods: ROESY, NOESY, TOCSY, ADEQUATE, and INADEQUATE [15]. Other methods can also be used by the system through a flexible procedure that allows input and processing of experimental 2D NMR data.

Prior to the structure generation the MCD is checked for the presence of contradictions (see Sect. 2.1.4).

The data collected in connectivity tables and graphically presented as an MCD are used as the input information for the 2D structure generator. If MCD checking shows that the 2D NMR data are consistent then Strict Structure Generation is initiated. Structures are generated under constraints determined from the molecular formula, the MCD, and any additional constraints which may be introduced by the chemist. The structure generator is based on mathematical algorithms developed by Molodtsov, who enhanced them during the Structure Elucidator program development [6, 14].

In general, structure generation is initiated by the command **Structure Elucidation/Run CSB Generator** (CSB, Correlation Spectroscopy Based). Structure generation will be performed from *all* existing MCDs in series upon executing this command. The command **Structure Elucidation/Run SCB Generator from Current MCD** allows structure generation to be performed from one MCD selected by the chemist. Figure 2.10 shows the dialog window associated with the CSB Generator Options.

The meanings of the majority of options are intuitively clear, but some of them require explanation. **Estimate Generation Time Only** is selected if the initial data is fairly uncertain and it is desirable to provide an estimate of the generation time and the number of generated structures is expected to be manageable.

The options **Add Generation Results to User Notes** and **Save Project After Generation is Completed** are recommended to be selected if it becomes clear that structure generation will be time-consuming (for instance, if the program is left to work overnight).

A group of options **Use Connectivities from** allows the chemist utilizing different combinations of 2D NMR spectra during structure generation. For instance, the presence of 3–5 nonstandard connectivities of 4J length in the COSY data may

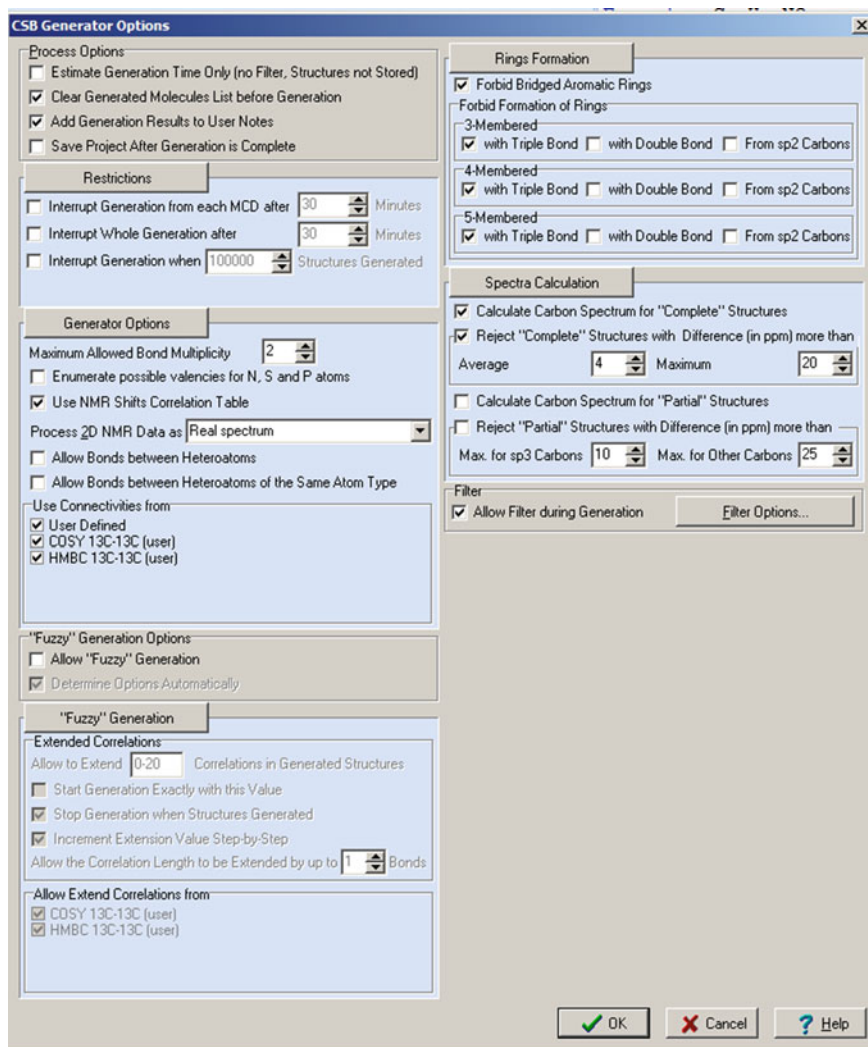


Fig. 2.10 CSB Generator Options

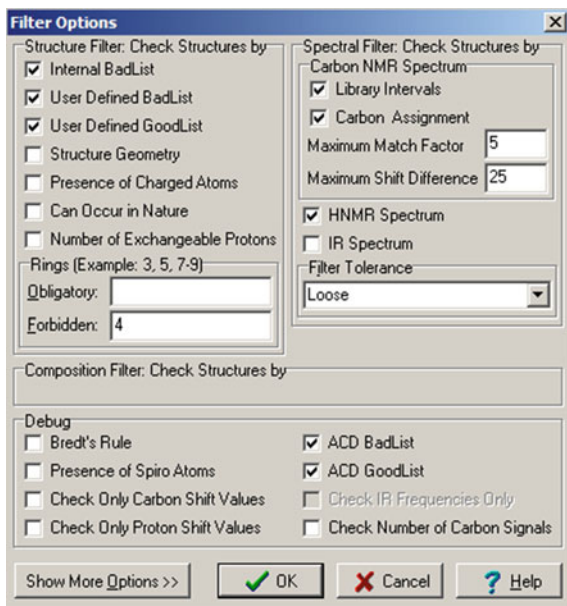
markedly elongate the time of FSG which makes it more advantageous to deselect the COSY check box and proceed with structure generation from the HMBC data only.

The options **Spectra Calculation** and **Allow Filtering during Generation** are used in combination and play an important role in the structure elucidation process. The purpose of these options is to reject generated structures for which predicted chemical shifts differ dramatically from the experimental shifts. On the level of the spectral filter, a *rough* spectrum prediction is realized by intervals of characteristic spectral features. If at least one ^{13}C or ^1H NMR chemical shift assigned to an atom

of a generated structure does not meet the corresponding spectral interval, the structure is rejected by the filter. As it is impossible to take into account the influence of all conceivable environments of a given atom in a molecule on an atom's chemical shift sometimes the filter can reject a correct structure. Therefore, to minimize risk of the correct structure loss the **Spectrum Calculation** option can be used. For this aim it is necessary to select the commands **Calculate Carbon Spectrum for Complete Structures** and **Reject Complete Structures with Difference (in ppm) more than** and indicate the average and maximum deviations which will be used as thresholds for the current structure rejection. ^{13}C chemical shift prediction is performed by the *Incremental* approach which shows a high speed of calculation ($\sim 30,000$ chemical shifts/s). In reality, it is practically impossible to notice if the ^{13}C chemical shift calculation was used or not used during the structure generation. Experience has shown that the thresholds $d = 4\text{--}5$ and $d_{\text{max}} = 20\text{--}25$ ppm are optimal values providing an output structural file of a manageable size. When the option **Spectrum Calculation** is used, in the dialog window **Filter Options** a check box **Carbon Assignment** is automatically selected and the fields **Maximum Match Factor** and **Maximum Shift Difference** are automatically filled in as shown in Fig. 2.11.

Thus, when ^{13}C spectrum calculation is activated the filter can be used both as a mechanism intended only for structure rejection in correspondence with the calculated ^{13}C NMR chemical shifts and as a “standalone” spectral and structural filter. If the filter is used *only* as a facility of the ^{13}C chemical shift calculation procedure then the check boxes **Library Intervals** and **HNMR Spectrum** must be deselected, while all structural constraints shown in the left part of the dialog window can be

Fig. 2.11 Dialog box **Filter Options**



used at the same time. If the filter is used in a mode where **Spectrum Calculation** is disabled it is necessary to check if the option **Carbon Assignment** is also deselected, otherwise the filter will reject all structures because they have no deviations for comparison. It is expected that if **Spectrum Calculation** and spectral filtering are used simultaneously (check boxes **Library Intervals** and **HNMR Spectrum** are selected) the output structure file will be of minimum size.

A question arises: Which method of reducing the output structural file and rejecting deliberately invalid structures is optimal? We suppose that the sequence of operation which was traditional for all expert systems “Fragment Selection → Structure Generation → Structure Spectral Filtering” can be modified now. The Structure Spectral Filtering can be replaced by Spectrum Calculation and utilization of the *spectral filtering* only for rejection structures which do not satisfy the threshold criteria $d \leq 4-5$ and $d_{\max} \leq 20-25$ ppm. As a result only structures that went through this stage will be saved. In this case, checking structures by ^{13}C and ^1H NMR characteristic spectral intervals can be used as an additional aid for verification of the saved structures.

Generated structures can be inspected by clicking on the key **MOL** in the toolbar of the main window of the program. Then ^{13}C NMR spectrum prediction is performed for all structures included in the output file and the structures are ranked in order of increasing d value using the method described in detail in the following section. To perform the NMR spectrum calculation it is necessary to press on the key **Tools** on the Toolbar and select a spectrum that should be predicted in the drop-down menu (Fig. 2.12).

If the user wants to predict chemical shifts for several kinds of NMR spectra by different methods of spectrum calculation, the command **All Spectra...** should be activated. In the dialog window **Select Spectra to Calculate** (Fig. 2.13) the corresponding spectra are chosen by check box selections. As a default chemical shifts will be calculated for all structures in a given structure file, but a **Maximum Number of Processed Structures** can also be specified.

If the user has his own structural hypothesis the proposed structure can be drawn in the **PM** (Proposed Molecule) window and ^{13}C and ^1H chemical shift assignment can be performed (use the tool **Edit Atom Properties**). Then the system can be used for verification of the proposed structure and the associated ^{13}C and ^1H NMR signal assignments by all available two-dimensional NMR spectra. For this purpose, a command **Structure/Check by 2D NMR Data (1)** is activated. If any nonstandard connectivities are found, the program displays a textual message detailing the cause(s) of the conflict(s). At the same time all connectivities are shown on the structure in graphical form. To ease visual analysis, nonstandard connectivities are marked in red.

2.2.1.1 Selection of Preferable Structure

As previously mentioned, in general, selection of the preferable structure is reduced to NMR chemical shift prediction for structures included into the output file and

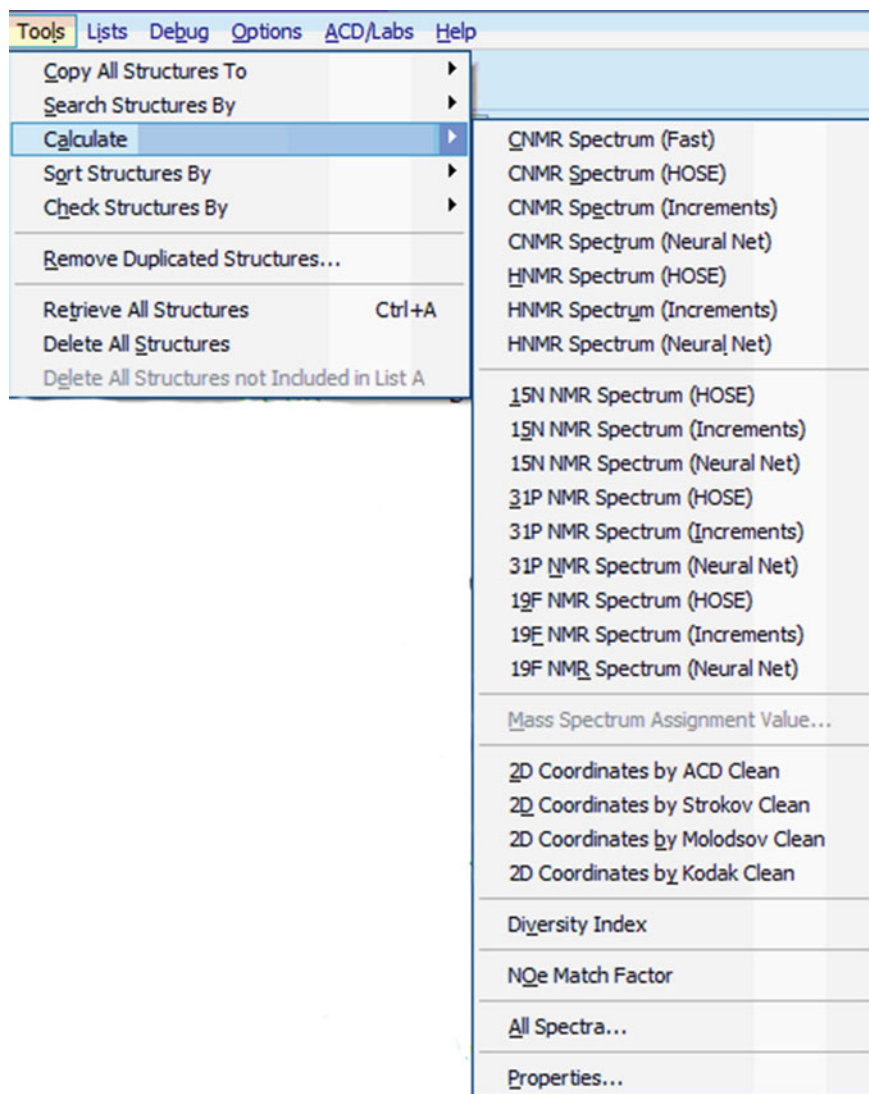
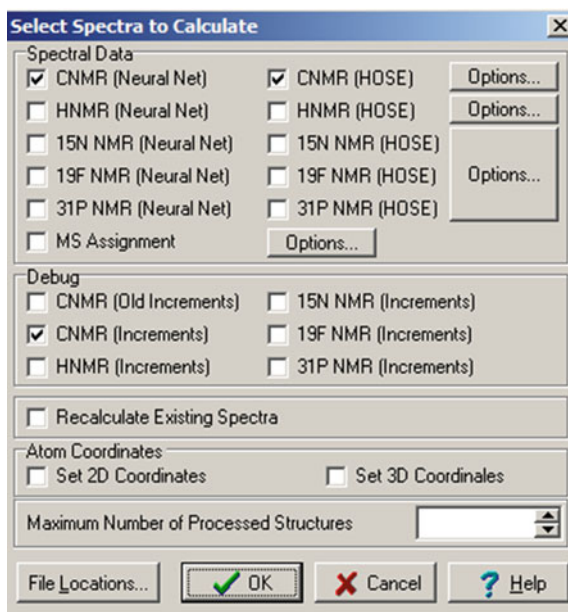


Fig. 2.12 Drop-down menu **Tools**

structure ranking in increasing order of deviation values. The StrucEluc system provides the following procedure for identifying the most probable structure in the output file.

First step ^{13}C NMR spectra are predicted for all generated structures using an incremental method, the fast method, and d_I values, the average deviation of an experimental ^{13}C NMR spectrum versus predicted chemical shifts, are calculated.

Fig. 2.13 The dialog window
Select Spectra to Calculate




If ^{13}C chemical shift calculation is used during structure generation this step is skipped.

Second step Duplicate structures in the output file are deleted (**Tools\Remove Duplicate Structures...**, Fig. 2.12). Among the generated structures there are usually duplicates that differ from each other only in terms of the assignment of the chemical shifts to different carbon atoms. If this possibility is not appropriately considered when deleting isomorphic structures, then the structure with the correct assignment of the chemical shifts could conceivably be the deleted isomorphic structure. To avoid this eventuality, the system executes a special procedure for duplicate removal. For each duplicate family only the structure that has the minimum d_I value is retained in the file as “the best representative” of the family. After duplicates are removed, the structures are then ranked by the d_I value and sorted in ascending order (**Tools\Sort Structures by\C NMR Spectrum (Increments)**, Fig. 2.12). The smallest d_I value indicates the best match between the experimental and calculated spectra and this structure will therefore be the first in the list. Experience shows that the incremental calculation of ^{13}C NMR spectra and their subsequent ranking usually places the correct structure among the first several structures at the top of the list. Only in rare instances will the correct structure be listed below fifth place. Such a preliminary ranking of the big resulting files can help to reject thousands and tens of thousands of structures that are known to be unsuitable.

Third step ^{13}C chemical shift prediction is carried out using an NN algorithm and the structure file is ranked again with d_N deviations. If the resulting file is extremely large the calculations can be applied only to the first several thousand

structures (it will take several seconds). As a result of this step the preferable structure is selected with greater reliability.

If the initial structure file is not too large it is desirable to perform removal of duplicates after calculating d_I and d_N values: selection of the same best structure by both incremental and neural net approaches will raise the probability of obtaining the right structure with correct chemical shift assignment.

For the user's convenience the ranked file can be displayed in a **Tile** mode for which it is necessary to press the icon  to **Display Structure and Tile** in the Toolbar.



A right mouse click on the field where the list of structures is displayed leads to the appearance of a menu (Fig. 2.14). This menu contains intuitively understandable functions which allow the user to manage information associated with structures displayed in the **Tile** mode.

Fourth step During the fourth stage ^{13}C NMR spectra are usually calculated for the first 10–50 (sometimes up to 100) structures of the ranked file using a fragmental method based on the HOSE code approach (this procedure can take several minutes). The average deviation values between the experimental and calculated values (d_A) are found and the structures are again rank ordered. Subsequent ranking increases the probability of moving the correct structure to the first position in the list. For

Fig. 2.14 The menu for managing information related to the structures which are displayed in the **Tile** mode

Switch to Table View	
Copy to Editor	ENTER
Zoom In	
Zoom Out	
Select Current	Ins
Select All	Ctrl+A
Unselect All	
Invert Selection	
Set List from Selection	
Save Selection...	
Load Selection...	
Select Data...	Gray *
Select Font...	
SS Search Color Options...	

additional control over the correct choice of the output structure, the HOSE code-based proton chemical shifts can be predicted and displayed together with the corresponding deviation value d_H . A complex match factor $d_{\text{complex}} = d_N(\text{C}) + 10d_N(\text{H})$ can also be used for ranking the structures in the output file.

The position of the correct structure in the file determines its rank depending on the type of ranking parameter, i.e., d_N , d_A , d_I , d_H , or d_{complex} correspondingly. The “rates” of the correct structure in the ranked file are denoted as r_N , r_A , r_I , and r_H . If the correct structure is the first in the list ranked by d_A values, then $r_A = 1$. As a rule, the final structural ranking is carried out according to increasing d_A and d_N values, while magnitudes of the d_I and d_H parameters serve as secondary aids for estimating the reliability of the correct structure selection. The accuracy of chemical shift prediction for each carbon atom can be evaluated visually by pressing the toolbar button **Show/Hide Carbon Assignment**. The accuracy is marked by colored circles on the atoms, while the following colors are used: *green*—the difference Δ between the experimental and predicted chemical shifts is not higher than 3 ppm ($\Delta < 3$ ppm), *yellow*— $3 < \Delta < 15$ ppm, *red*— $\Delta > 15$ ppm. All kinds of information related to structures can be visualized using the following icons on the toolbar:  and . The first of

them allows the experimental and predicted ^1H , ^{13}C , and ^{15}N NMR chemical shifts to be displayed as well as different representations of atom numbering (Fig. 2.15, left). The second button is used to display the kinds of connectivities that are selected by the user (Fig. 2.15, right).

The top structures or selected structures displayed in the **Tile** mode can be copied to the ChemSketch window by the command **File>Create Report/List of Structures** as shown in Fig. 2.16.

When the first and second ranked structures contain markedly differing structural elements, then the prediction of the MS match factor (m_i , where i is the position of a structure in the ranked file) may also be useful for confirmation of the preferable structure. For this purpose, it is necessary to activate the command **Mass Spectrum Assignment Value**, Fig. 2.12. The system utilizes a routine that is capable of calculating the percentage of peaks in the experimental MS spectrum that can be interpreted on the basis of a given structure. The calculation of the MS match factor is relatively time-consuming, so it is worth using it only in those cases when the difference $\Delta_{(2-1)} = d_A(2) - d_A(1)$ is small. Here $d_A(1)$ and $d_A(2)$ represent the deviations corresponding to the first and second structures in the ranked file.

In ambiguous cases it may be useful to display the calculated ^1H NMR spectra in graphical form. Also, to facilitate structure analysis in the output file, the *StrucEluc* system is supplied with a feature that calculates structural similarity coefficients (**Structure/Similarity Search in/Generated Molecules**). In this way if the investigator has an idea of the class of structure under investigation he can use this structure as an input to allow rank ordering relative to the structural similarity of the results file.

Numeration	
$A_{\#}$	Show Atom Numbers
^{13}C	Show Experimental Labels
^{13}C	Show Experimental Numbers
1H	Show Experimental Labels
1H	Show Experimental Numbers

^{13}C NMR Peak Values	
<input checked="" type="checkbox"/> ^{13}C	Experimental
<input checked="" type="checkbox"/> ^{13}C	Calculated (HOSE)
<input type="checkbox"/> ^{13}C	Calculated (Fast)
<input checked="" type="checkbox"/> ^{13}C	Calculated (Increments)
<input checked="" type="checkbox"/> ^{13}C	Calculated (Neural Net)
<input type="checkbox"/> ^{13}C	Calculated (QM DFT)

1H NMR Peak Values	
<input type="checkbox"/> 1H	Experimental
<input type="checkbox"/> 1H	Calculated (HOSE)
<input type="checkbox"/> 1H	Calculated (Increments)
<input checked="" type="checkbox"/> 1H	Calculated (Neural Net)

^{15}N NMR Peak Values	
<input type="checkbox"/> ^{15}N	Experimental
<input type="checkbox"/> ^{15}N	Calculated (HOSE)
<input type="checkbox"/> ^{15}N	Calculated (Increments)
<input checked="" type="checkbox"/> ^{15}N	Calculated (Neural Net)

^{17}O NMR Peak Values	
<input type="checkbox"/> ^{17}O	Experimental

Filter by Nuclei	
Show For Active Spectrum	
<input checked="" type="checkbox"/>	1H - 1H Correlations
<input checked="" type="checkbox"/>	^{13}C - 1H Correlations
<input checked="" type="checkbox"/>	^{15}N - 1H Correlations
<input checked="" type="checkbox"/>	Other Correlations

Filter by Bond Distance	
<input checked="" type="checkbox"/>	2J Correlations
<input checked="" type="checkbox"/>	3J Correlations
<input checked="" type="checkbox"/>	4J and More

Filter by Correlation Type	
<input checked="" type="checkbox"/>	Through Bonds
<input type="checkbox"/>	Through Space

Geminal Coupling	
Show Geminal 1H - 1H Couplings	

Ambiguous Correlations	
<input checked="" type="checkbox"/>	Show Outranged Ambiguous Correlations

Fig. 2.15 *Left* visualizing experimental and predicted 1H , ^{13}C , and ^{15}N NMR chemical shifts and different types of atom numbering. *Right* visualizing connectivities of different origin and different length

When the best structure is selected the following question can be posed: which structures would be generated if all theoretically possible HMBC and/or COSY correlations (types of 2D NMR spectra are selected by the user) were observed in the experimental spectra? The command **Structure>Create Project from Structure** is used to provide the answer to this question. The program creates a new

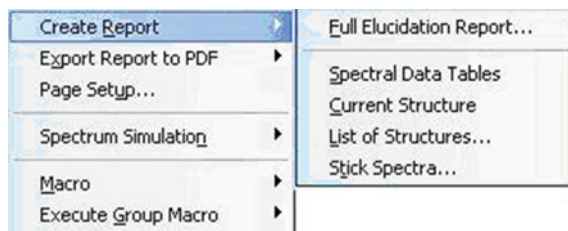


Fig. 2.16 A part of the **File** menu containing commands which are utilized for copying a List of Structures to ChemSketch

project with the MCD containing all theoretically possible connectivities from which structure generation can be performed.

2.2.2 Application of Fragments in Combination with 2D NMR Data

Computer-assisted structure elucidation using 2D NMR data is quite efficient for the elucidation of structures of complex organic molecules. However, if the structural restrictions imposed by the MCD are not sufficient for the generation of a reasonable number of possible structures within an appropriate time, it is to be expected that the utilization of molecular fragments can help facilitate the solving of the problem. Commonly appropriate fragments to aid in the solution of a problem can be found in the Structure Elucidator knowledge base. *The main advantage of these fragments is that all fragment carbon atoms are supplied with the ^{13}C NMR assignments obtained from the full structures that were used for creation of the fragment database-Fragment Library.*

The first step of the process is a fragment search of the Fragment Library which is initiated by the command **Structure Elucidation\Search Fragments by C NMR Spectrum** (Fig. 2.17). To change the default options the researcher can use the dialog window **Search Fragments by ^{13}C NMR Spectrum Options** (Fig. 2.17).

If the radio button **From Molecular Formula** is selected, all the Found Fragments will have molecular compositions not exceeding the Molecular Formula. Utilization of 2D NMR connectivities and usage of the filter during fragment search can optionally be activated by the user. When the fragment search process is completed the Found Fragments are displayed in the **Found Fragment** window (**View\Structures List\Found Fragments**, Fig. 2.18).

As a result of the fragment search a set of L Found Fragments is selected. The next step is the creation of MCDs using the found fragments (FF), for which the command **Structure Elucidation\Create MCDs Using Fragments** is activated (see Fig. 2.6), and then the dialog window **Create MCDs Options** is opened (Fig. 2.19). For the first run automatic determination of most options is allowed.

Fig. 2.17 Dialog window
Search Fragments by ^{13}C
NMR Spectrum Options

Search Fragments by ^{13}C NMR Spectrum Options

Search Options

☒ Clear Found Fragments before Search

Spectral Data

☒ Reject structures with Match Factor more than 5 ppm

☐ Reject structures with number of signals less than 20 % in spectrum

☐ Allow lack of signals in "full" structures: 5 signals

☐ Allow excess of signals in structures: 2 signals

☐ Allow excess only for quaternary carbons

☐ Ignore peak intensity during Search

Composition

☒ Check composition

☒ From Molecular Formula ☐ From Defined Composition

Composition: C(0-12) H(0-7) O(0-2) Cl(0-1)

Monoisotopic Mass

☒ Check Monoisotopic Mass

Monoisotopic Mass: 0.000-218.513 Tolerance (Da) 0.5

2D NMR Data

☒ Use HSQC information (check: chemical shifts of attached hydrogens)

☐ Use HMBC and COSY information (check: distance between chemical shifts)

Tolerances

	^{13}C	^1H	
Tolerance for "First Sphere" Atoms (no less than)	12	2	ppm
Tolerance for "Second Sphere" Atoms (no less than)	6	1	ppm
Minimum Possible Tolerance for All Atoms (no less than)	3	1	ppm
Maximum Possible Tolerance for All Atoms (no more than)	20	2	ppm

Filter

☐ Allow Filter during Search Filter Options...

Search Databases

☒ ACD Internal Full Structures Database

☒ ACD Internal Fragments Database

Internal DBs

Add...

Remove

Up

Down

Spectral Data... OK Cancel Help

For MCD creation the selected number of FFs can be set either by the operator or by the program—automatically. The main idea of the algorithm that implements this procedure is as follows. The chemist defines the number of fragments, l ($l \leq L$), that will be used for MCD creation and sets an error, E , that defines the maximum difference allowed between the chemical shifts of the fragment carbons and the corresponding values observed in the experimental spectrum under study. The situation is common when several experimental chemical shifts are close to the

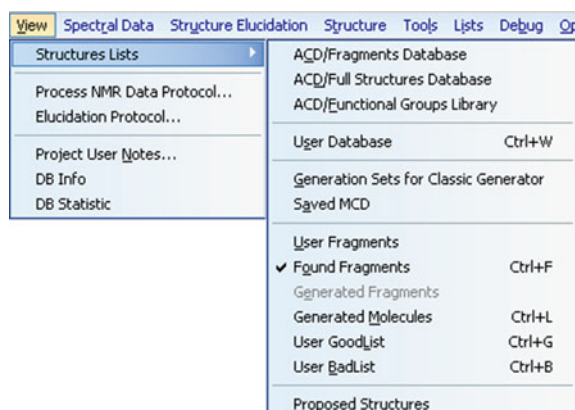


Fig. 2.18 A part of the **View** menu used for switching windows containing structural files common for Structure Elucidator

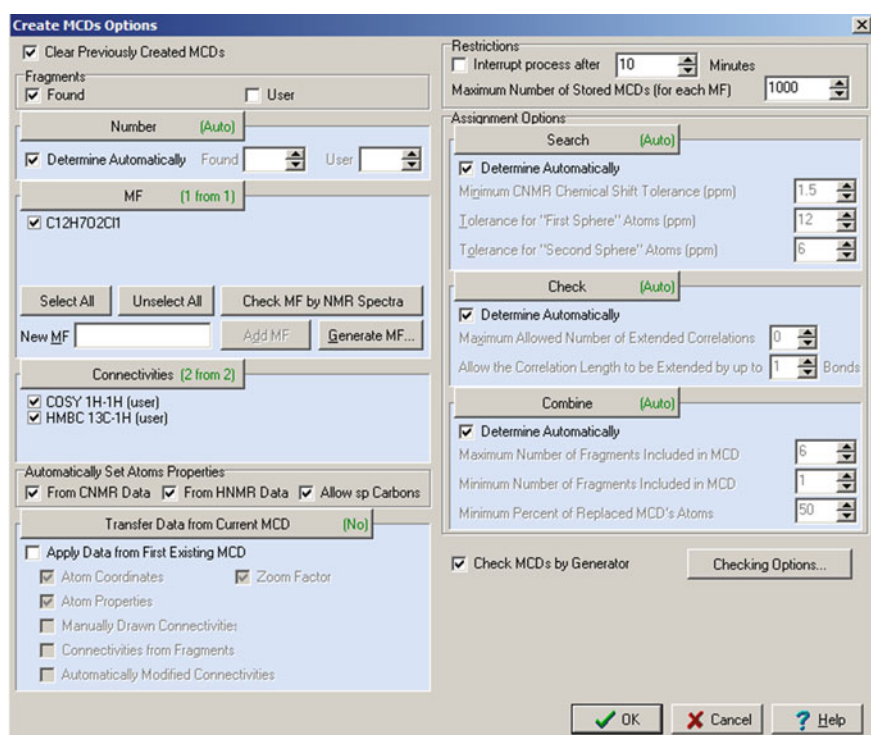


Fig. 2.19 The dialog box **Create MCDs Options**

chemical shift assigned to a given carbon atom of a fragment. It is important to note that both parameters, l and E , are closely interrelated and choosing the most efficient values may be a matter of trial and error.

The ^{13}C NMR subspectrum of each fragment is compared with all experimental chemical shifts. The number of hydrogen atoms attached to a carbon atom is taken into account during this process. Consider a fragment that contains n carbon atoms and an arbitrary atom C_i of the fragment has a chemical shift δ_i ($i = 1 \div n$) and multiplicity m_i . Suppose that the experimental chemical shifts $\delta_{i1}, \delta_{i2}, \dots, \delta_{iq}, \dots, \delta_{ip}$ meet the conditions $|\delta_i - \delta_{iq}| \leq E$ and $m_i = m_{iq}$. Then, all possibilities of substituting the δ_i values for the experimental values $\delta_{i1}, \delta_{i2}, \dots, \delta_{iq}, \dots, \delta_{ip}$ must be verified.

If the conditions $|\delta_i - \delta_{iq}| \leq E$ and $m_i = m_{iq}$ hold for all f carbon atoms, then the given fragment is recognized as a candidate for inclusion in the process of creating the MCD. If this condition does not hold then the fragment is excluded from consideration. The program also checks whether the carbon atom assignments correspond to the experimental chemical shift correlations comprising the skeletal atoms making up the fragment. The fragments that survive the test are then included in the set of *prospective* fragments.

The more the skeletal atoms “absorbed” by the fragments, the shorter is the process of structure elucidation. With this in mind an algorithm that combines the prospective fragments within one MCD was developed. To realize this procedure, all possible combinations of prospective fragments are searched and only combinations that are in agreement with the experimental 2D NMR correlations are chosen. The fragment combinations that pass this examination form a set of prospective fragment combinations. These fragments are then “projected” onto the MCDs together with any remaining free atoms. The user can then visually analyze these MCDs.

The total number of MCDs, n_{MCD} , depends on the following parameters which are defined by the user:

- L number of found fragments which will be used for the creation of MCDs ($l \leq L$);
- n_f the minimal number of fragments that must be present in each MCD;
- q the minimum percentage of all skeletal atoms that must be absorbed by the fragments present in each MCD.

In general the more the atoms that are “absorbed” by the fragments accepted by an MCD, the greater the likelihood that the process of structure generation from the given MCD will be more time efficient.

The speed of structure generation depends on the size of the molecular fragments. If the number of small fragments composing the MCD is large enough, then this will speed up the generation. Structure generation is much faster when the MCD comprises a small number of big fragments. Depending on the size of the molecule being analyzed and the size of fragments placed at the beginning of the ranked list of FFs, the n_f value is usually defined as a number from 1 to 4. The most efficient results are obtained if q is significant, generally 40–60 %. The dialog window which is used for setting the minimum and maximum numbers of fragments included into the created MCDs is presented in Fig. 2.20.

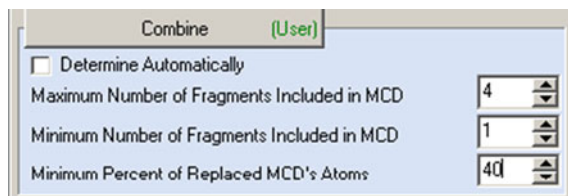


Fig. 2.20 The dialog window **Combine** which is used for setting the minimum and maximum number of fragments included into the created MCDs

The conclusion of all further verification procedures is a check of all produced MCDs for contradictions. The program offers an option that deletes all MCDs that are recognized as contradictory. The diagrams remaining after checking can be used in the structure generation process. The user has the opportunity to omit the connectivity verification because contradictory MCDs will be detected and rejected in the process of structure generation. Moreover, for the process of structure generation the user can select one or more MCDs that are attractive to the user who may have prior knowledge of a particular structure class or target structure. To alleviate having to choose a preferable MCD they are automatically ranked in order of the increasing number of free carbons. In this way it is possible to select a series of appropriate MCDs, starting from that ranked first.

The number of MCDs produced from a given set of fragments can be rather large (sometimes the n_{MCD} value is greater than one thousand). To provide for the possibility to edit a big set of MCDs a special procedure was elaborated, which allows one to transfer all changes made in the *first* MCD to all the MCD sets (**MCD Apply Properties to All MCDs**). In particular, it is possible to specify options which transfer atom coordinates, zoom factor, atom properties, manually drawn connectivities, and connectivities automatically modified during the MCD checking for presence of contradictions. This procedure essentially alleviates using a priori information in the fragment mode of structure elucidation.

In the process of analyzing a novel compound it is entirely possible that there will be no fragments in the database that will reduce the magnitude of the challenge. It is natural in such cases to expect that the introduction of user-defined fragments may help to form the MCDs. The main qualitative difference between a found fragment (FF), and a user fragment (UF), is that the FF already contains carbon atoms with assigned chemical shifts while the carbon atoms of the UF have no carbon chemical shift assignments. Two ways have been suggested to introduce UFs into the program:

- Calculate the carbon chemical shifts of the fragment using the HOSE code-based method (see Sect. 1.4.1.1);
- Search the KB for fragments that *comprise* the user fragment.

It is likely that fragments from at least one of the two sources would be available for use by the program.

2.2.2.1 Choice of E Value

In the process of MCD creation from fragments, the E value is of great importance since it markedly influences the result of applying the fragments. There are a number of principles governing the selection of the E value. As a rule, the smaller the value of E , the smaller the number of MCDs, n_{MCD} , created from FFs. The advantage of a small number of MCDs is of course that it can reduce the time for structure generation, t_g . At the same time, t_g is also a function of the *fragment dimensions*. Larger fragments generally shorten the structure generation process. However, if a fragment is large and correspondingly contains many assigned carbon atoms, then as a consequence it is not as likely that *all* carbon atoms, especially the terminal ones, of a large fragment will fit the experimental shifts thereby satisfying a narrow interval for $\pm E$. The program automatically sets the E value for terminal atoms equal to 12 ppm to account for this issue (Fig. 2.21).

Large fragments are the most useful but to utilize them in the structure elucidation process a large E value is frequently necessary. A large E value can correspondingly increase the n_{MCD} value. The optimal approach would be to set a large enough E value and select only those MCDs containing large fragments for the structure generation. This principle therefore justifies manual (user) or automatic rejection of MCDs containing small fragments. With testing it has been shown that the optimal program parameter controlling the minimum number of carbon atoms in the fragment used for MCD creating should be set to a value of five. Unfortunately, it is impossible to determine an optimal E that is valid for a diverse range of problems. The value of E should be optimized for each task by gradually increasing the E value starting from 1.5 ppm. This procedure can be performed manually or automatically during structure generation.

2.3 Nonstandard Correlations and Fuzzy Structure Generation

2.3.1 Challenge of Nonstandard Spectral Responses

CASE 2D NMR methodology can provide solutions for computer-assisted structure elucidation tasks in a reasonable time if the initial data (NMR, MS, IR, chemical

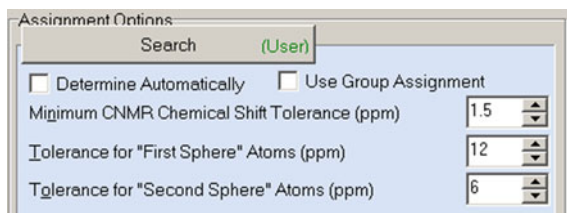


Fig. 2.21 Selection of ^{13}C chemical shift tolerances for creation of MCDs from FFs. Minimum C NMR chemical shift tolerance E is 1.5 ppm

assumptions, etc.) are *true, consistent, and complete*. The latter means that the number of observed 2D NMR correlations is large enough to sufficiently define the connectivities within a structure. If at least one of these conditions is violated the possibility to somehow find a correct solution to the problem decreases significantly. As mentioned earlier, methods to overcome the presence of contradictions in experimental data have been suggested. The corresponding algorithms have been developed and implemented in the StrucEluc system. In this chapter we will consider the different StrucEluc-based strategies for molecular structure elucidation in those situations when 2D NMR spectra contain nonstandard correlations (NSCs).

As a result of a series of computational experiments it has been shown that the program was capable of determining the *presence* of connectivities of nonstandard length in 90 % of all cases using the MCD checking procedure described in Sect. 2.1.4. These results are very encouraging since routine experimental methods guaranteeing the precise determination of COSY and HMBC connectivity lengths are not available. Knowledge of the presence of contradictions in 2D NMR data gives the investigator valuable information that can determine the strategy of structure elucidation with these data. An erroneous program report regarding the presence of nonstandard connectivities may appear if properties of at least one atom are assigned incorrectly (for instance, label “*fb*” is set instead of “*ob*”). A “false” message can also appear in those relatively rare cases when an unknown under study contains a pair of bonded heteroatoms, but the absence of such atomic pairs is set in the program options. In these situations a program message regarding the existence of contradictions can help the chemist to reveal the presence of bonded heteroatoms (see examples in Part III). There were no other cases of incorrect detection of nonstandard connectivities where contradictions in 2D NMR data were not present. The program frequently not only identifies the contradictions in the data correctly, but is able to successfully remove them automatically to allow determination of the correct structure. The program was unable to detect the presence of NSCs when 2D NMR data mainly contained only one HMBC nonstandard connectivity. This occurrence can be explained by the fact that if there are only one or two HMBC nonstandard connectivities in the data, the atoms in a conceivable structure may be arranged so that their arrangement complies with the standard length of all connectivities. If the number of NSCs is large, such an arrangement of atoms is unlikely. The presence of *implicit* nonstandard connectivities can become apparent as a result of structure generation and subsequent structure filtration with the use of spectral libraries: if all the generated structures obviously contradict the spectral data, the program produces an empty results file. Indirect evidence of the possibility that contradictions were not detected may not only be an empty result file but large values, more than 3.5–4.5 ppm, of the chemical shift deviations, d_A and d_N , calculated for the first ranked structure. Investigations have shown that nonstandard connectivities were detected by both direct and indirect methods for 95 % of the analyzed tasks containing contradictory data. If there are reasons to assume that the program did not detect contradictions in the initial data it would be highly likely that the problem could be solved with the use of Fuzzy Structure Generation as will be described in Chap. 5.

Since it is possible that 2D NMR data can contain implicit nonstandard connectivities, the most probable structure generally requires additional verification by independent methods. Particularly, incorrect structures can be rejected on the basis of predicted chemical shifts and multiplets in the ^1H NMR spectrum. However the most effective method, as we will see, is application of FSG. If the structure is generated after automated removal of contradictions then it is still desirable to check for the presence of nonstandard connectivities. The connectivities can be verified with appropriate experimental parameter optimization to probe the values of the spin couplings [4, 5].

Therefore, it is not always possible to find nonstandard connectivities and to automatically resolve the contradictions in 2D NMR data sets. In practice, the following difficult situations accounting for the presence of NSCs may typically arise:

1. The program detects the presence of NSCs and makes an attempt to remove the contradictions in the data but then reports that contradictions cannot be removed automatically. Frequently, FSG can help to solve the problem. Generally, additional experiments are required in an effort to detect NSCs.
2. The program fails to detect nonstandard connectivities and displays a message informing the researcher about the absence of contradictions. In this case strict structure generation is initiated. The following outcomes are possible: (a) no structure is generated and saved after filtering; (b) the wrong structure(s) is generated, which can generally be recognized because of the large values of the ^{13}C experimental versus predicted deviations. Again FSG can help in this situation.
3. The program detects the presence of nonstandard connectivities, makes an attempt to remove the contradictions in the data, and displays a message that the contradictions were removed, though in fact, some contradictions still remain. This is due to the fact that not all nonstandard connectivities are lengthened. Possible undesirable consequences, and the methods to overcome them, are similar to those listed for Point 2 above.

It should be obvious that the most dangerous situations are when incorrect solutions are produced and these can occur for Points 2 and 3 above. Even in those cases when the program is not able to remove detected contradictions, specifically case 1, the fact that contradictions are detected is of great importance for selection of an optimal method of problem solving.

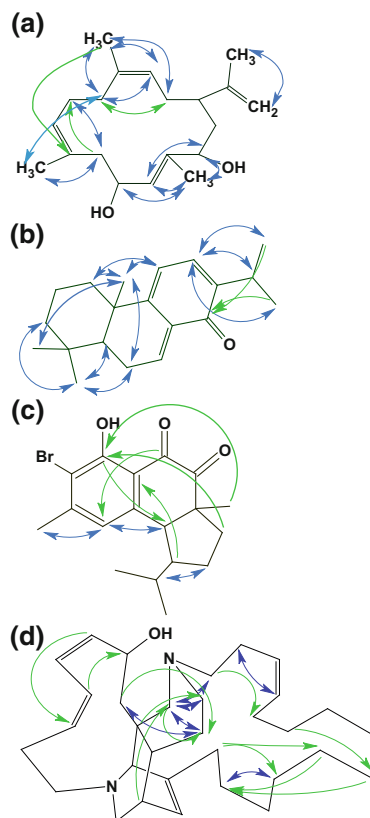
2.3.2 Solving Problems Using Fuzzy Structure Generation

In those cases when correlations are present in the 2D NMR data with nJ where $n > 4$, the method of automatic removal of contradictions, unfortunately, does not work. The augmentation of the path between two intervening nuclei by one bond obviously cannot lead to the generation of a correct structure in this case. Moreover,

due to a lack of constraints that are to be logically analyzed, even in those cases when $n = 4$ the algorithm gives no guarantee that all nonstandard correlations will be found and corrected. For example, the greater the number of carbon atoms with properly defined properties (in regard to the type of hybridization and different heteroatom neighborhoods) and/or the higher the total number of available 2D NMR connectivities, the higher the probability of successfully performing logical analysis to arrive at the correct structure. In contrast, severely proton-deficient molecules are among the most challenging. Obviously, the problem becomes more computationally complicated as the size and complexity of a molecule increases.

The number of NSCs contained within the 2D NMR data associated with a molecule, m , can be rather large—up to about 20 correlations. At the same time, to remove contradictions the augmentation of standard correlation lengths, a , could be 1–3. As an example of such situations several structures taken from the literature are used to demonstrate those examples with a large number of NSCs including 5J and 6J coupling constants (see Fig. 2.22).

Fig. 2.22 An illustration of a number of structures containing multiple nonstandard correlations.
a $m = 15$, $a = 3$ [16];
b $m = 13$, $a = 3$ [17]; **c** $m = 8$,
 $a = 3$ [18]; **d** $m = 18$, $a = 2$ [19]



The nonstandard COSY correlations are shown as blue arrows and the HMBC correlations by green arrows. In the legends for the structures m is the total number of nonstandard correlations, and a (augmentation) is the value of correlation lengthening allowed during the process of FSG.

To overcome the described difficulties, a computational approach was suggested that has been defined as Fuzzy Structure Generation.

2.3.3 Modes of Fuzzy Structure Generation

Numerous computational experiments have allowed us to conclude that if the program detects the presence of NSCs but fails to resolve contradictions in the 2D NMR data using algorithms described in Sect. 2.1.4, then FSG should be used to solve the problem. Moreover, it is quite probable that structure elucidation from 2D NMR data on the basis of FSG can be considered as a general CASE strategy because it is almost independent of the presence or absence of NSCs in the 2D NMR data.

FSG can easily be controlled by parameters that make up a set of options. The two main parameters are: m —number of nonstandard connectivities and a —the number of bonds by which some connectivity lengths should be augmented. Unfortunately, 2D NMR spectral data cannot deliver definitive information regarding the values of these variables and, as a matter of fact, both of them can be determined only during the process of structure elucidation. It has been concluded that in many cases the risk of choosing an erroneous value for a can be avoided and the solution of a problem can be considerably simplified if the lengthening of the m connectivities is replaced by their *deletion*. When set in the options the program can ignore by deletion connectivity responses that have to be augmented (by convention, the parameter a is set to a value of 16 in these cases). Such an approach can be successful in those cases when the number of 2D NMR connectivities is in some sense optimal. In this sense we mean that the total number of connectivities (structural constraints), N , must be large enough to facilitate description of the chemical structure. In many instances there are sufficient numbers of correlations in the ensemble of 2D NMR data acquired to essentially over determine the structure—in other words there is redundancy in some of the connectivity information. It can then be expected that deletion of m of the connectivities will not dramatically influence either the generation time or the size of the output file. On the other hand, the number of combinations of N connectivities taken m at a time can be very large. This can dramatically impede problem solving to a point that it is not feasible to solve the problem. Indeed, some researchers have commented that some of the ACCORDION-optimized long-range heteronuclear shift correlation experiments [15] actually provide too many long-range correlations of the type ${}^nJ_{\text{XH}}$ where $n \geq 4$.

If the total number of connectivities, N , is small then further decreasing N by m in a connectivity combination can lead to an excessive decrease in the number of structural constraints required for solving the problem. In such a case the problem

may be difficult to solve because the 2D NMR data structural constraints will only reduce the total number of possible isomers very slightly.

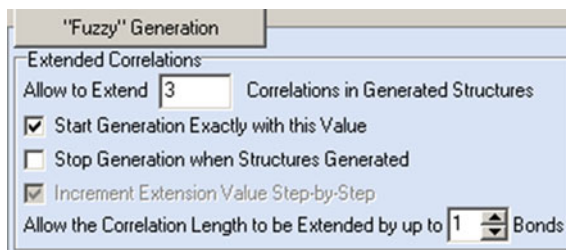
Independent of the use of augmentation or removal of connectivities, the crucial point in application of FSG is the number of connectivity combinations that should be checked during structure generation. For instance, if $N = 60$ and $m = 5$ then the number of connectivity combinations, $n_{\text{math}} = C_N^m$, is equal to ~ 5.5 million. Any attempt at structure generation has to be performed using each of these combinations. It is necessary to perform generation of structures from each of the C_N^m data sets and obtain the output file as a unification of all of the intermediate results. Although the StrucEluc structure generator is fast, the productivity is certainly insufficient in terms of coping with a combinatorial problem as outlined here.

To overcome this difficulty the system is delivered with an algorithm capable of reducing the number of combinations without the risk of losing the correct solution. The first step is to reduce the total number of connectivities N down to N_0 , where N_0 is the number of connectivities used to form the connectivity combinations. The data are preprocessed according to the following rules: (1) ambiguous connectivities are excluded from consideration; (2) if two connectivities C1 to C2 and C2 to C1 are present then only one of them is included in a data set. One of the two equivalent correlations is redundant and corresponds to overdetermination of the data needed for solution of the structure. The second and most important step is based on the results of logical analysis of the initial 2D NMR data. If connectivity sets containing NSCs are identified, then groups of these connectivities are utilized to produce connectivity combinations. As a consequence connectivities that are suspected to be nonstandard are included in all resulting combinations and the initial number of combinations reduces. In addition, the algorithm is capable of immediately detecting combinations of connectivities from which structure generation is impossible—a connectivity combination of this kind still contains at least one NSC. These combinations are skipped during the structure generation process. As a result FSG can be performed in a reasonable time even in those cases when n_{math} is very large. If the MCD checking process fails to detect nonstandard correlations in the 2D NMR data (the probability of failure is about 10 %) the program is forced to try all C_N^m connectivity combinations. This can drastically increase the time to solve the problem and the described approach is inefficient. In these cases, UFs and FFs can frequently be helpful. The ability of the program to calculate and display the real number of connectivity combinations to be validated during FSG allows the user to approximately evaluate the complexity of a given task even at the first stage of the structure elucidation process.

When option parameters are combined in a different way it is possible to initiate the following most appropriate modes of FSG:

Mode 1 Structures are generated such that the number of correlations that are extended is specified ($m = m_0$) and connectivity augmentation is also assigned ($a = a_0$). In this case for an HMBC correlation having a length of 1–2 skeletal bonds both the lower and upper length limits are updated and the connectivity length is extended to three bonds. Example: $m = 3$, $a = 1$, (Fig. 2.23).

Fig. 2.23 *Mode 1*: Example of fuzzy generation options



Mode 2 Structure generation is performed using the following options: it is assumed that the number of extendable (or ignored) connectivities can not exceed m_{\max} , ($m = 1, 2, \dots m_{\max}$), while a is equal to a_0 . The m_{\max} value is defined as the *maximum* allowed number of nonstandard correlations in the 2D NMR data. Typically the m_{\max} value is set equal to 20 thereby covering a wide range of nonstandard connectivities (see Fig. 2.22). The program initially performs structure generation with a value of $m = 1$. If the attempt is unsuccessful then the m value is *automatically* incremented by 1 and a new run is made with $m = 2$ and so on. An iteration is declared unsuccessful if either no structure is stored after structure generation and spectral filtration or if an *unacceptable* solution was found. When m reaches the m_g value the program considers the 2D NMR data to be consistent, then FSG is initiated with $m = m_g$. The program stops after completing structure generation with $m = m_g$ if the output structure file is not empty and if an *acceptable* solution is provided. Example: $m = 1-20$, $a = 1$, (Fig. 2.24).

Mode 3 The number of connectivities m is allowed to vary between m_{\min} and m_{\max} values ($m_{\min} \leq m \leq m_{\max}$), while the fixed number of bonds a_0 is set. The minimum number m_{\min} is usually derived as a result of checking the 2D NMR data for consistency. The program stops when similar conditions as described for *Mode 2* are achieved. Example: $m = 1-20$, $a = 1$, (Fig. 2.25).

Mode 4 This mode is a generalization of *Mode 3* where the interval for m value variation is defined by the condition $m_{\min} \leq m \leq m_{\max}$ at $m_{\min} = 0$. The peculiarity of this mode is that it is a “generalized” mode of structure generation and can be initiated with $m = 0$. In this mode, the program starts by checking the hypothesis that NSCs are absent in a given 2D NMR dataset. If the dataset does not contain nonstandard connectivities then the program completes the process of structure generation and the further solution of the problem is carried out as described previously (Sect. 2.2).

Fig. 2.24 *Mode 2*: Example of fuzzy generation options

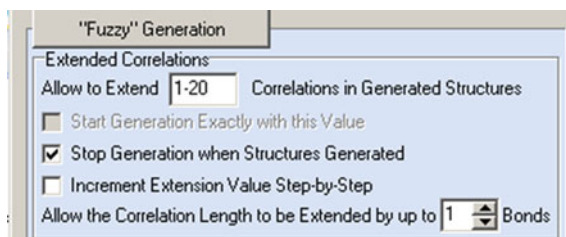
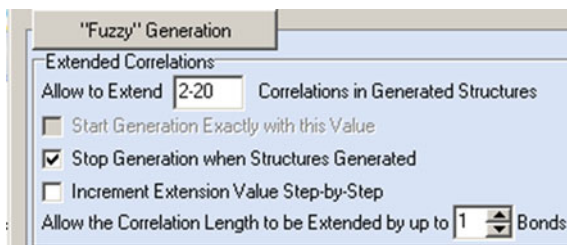


Fig. 2.25 *Mode 3*: Example of fuzzy generation options



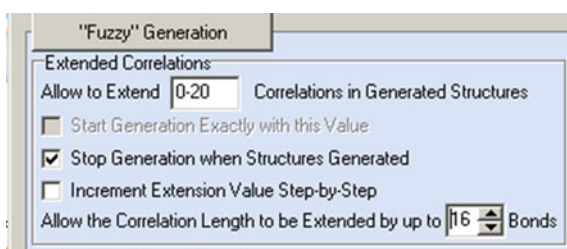
If an attempt with $m = 0$ proves to be unsuccessful then the program automatically performs FSG starting with $m = 1$, $a = 16$ and continues problem solving in the manner described earlier for *Mode 3*. The merit of such an approach is that no assumption regarding the a value is necessary. Example: $m = 0-20$, $a = 16$, (Fig. 2.26).

Mode 5 This mode is initiated if it is necessary to perform FSG iteratively covering all values of m starting from m_{\min} to m_{\max} without exclusion. For example, if structure generation is successful at $m = m_g$ then the program automatically switches to $m = m_g + 1$ and so on until it reaches $m = m_{\max}$. The structures generated at each step are added to those generated during the previous step. This mode is useful to check the solution for stability to make sure that the best structures found at steps $m = m_g$ and $m = m_g + 1$ or higher are equivalent. Examples: $m = 1-4$, $a = 16$:

- Generation is performed at $m = 1, 2, 3, 4$, and even if no structure is saved at some m value the generation will be continued at $m + 1$ and so on (Fig. 2.27).
- If no structure is saved at some m value the generation will be stopped (Fig. 2.28).

Mode 6 This mode resembles *Mode 5*, but the function of this mode is to generate all structures for which the number of nonstandard connectivities is *less or equal to* m at the given value of a . The corresponding options are denoted as $\{m \leq m_0, a = a_0\}$. The number of connectivity combinations from which FSG is performed depends only on the N_0 and m values. In contrast to the "step-by-step" modes some combinations of the connectivities are united by this approach and this in principle can speed up the calculations. When this procedure is performed only the maximal lengths of HMBC connectivities (i.e., two skeletal bond lengths) are

Fig. 2.26 *Mode 4*: Example of fuzzy generation options



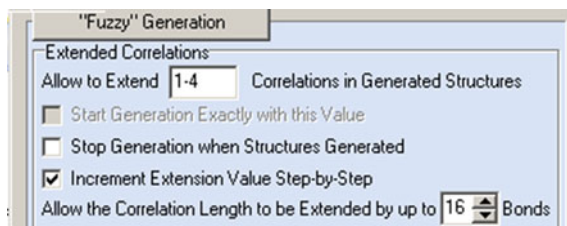
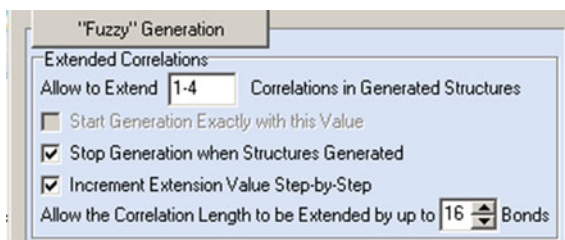


Fig. 2.27 Mode 5: Example of fuzzy generation options

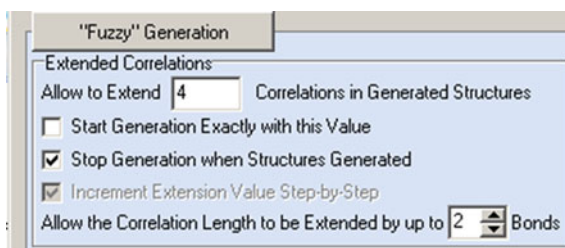
Fig. 2.28 Mode 5: Example of fuzzy generation options



enlarged. For example, consider an HMBC connectivity between C-1 and C-2 atoms whose “standard” length is varied from 1 to 2 skeletal bonds. In this mode the updated connectivity length varies from 1 to 3 skeletal bonds. It is important to note that the number of nonisomorphic structures generated in this mode is equal to the total number of nonisomorphic structures generated during all steps of *Mode 5*. However, the total time necessary for completion of FSG can be significantly different between these modes. Example: $m = 4$, $a = 2$, (Fig. 2.29).

Mode 7 This approach gives the researcher a chance to solve a problem in a fully automated mode. To initiate this mode the commands “**Allow Fuzzy Structure Generation**” and “**Determine Options Automatically**” are selected (see Fig. 2.10). The program analyzes the 2D NMR data and depending on the results makes a corresponding decision on the choice of the generation parameters and the

Fig. 2.29 Mode 6: Example of fuzzy generation options



strategy of their application. If a problem can be solved in the **Common Mode** (without using fragments) FSG with automatically determined parameters is very effective.

Note that *Mode 4* with the parameter *a* set equal to 16 (see an example below) can be considered as the most comprehensive mode since in principle it will solve a problem in which the 2D NMR data contain an *unknown* number of nonstandard connectivities of an *unknown* length.

If the problem is successfully solved with a given set of options then the real *m* and *a* values are reported by the program. Nonstandard connectivities are observed visually from the resulting structure which is displayed along with all COSY and HMBC connectivities. Nonstandard connectivities are easily recognized as they are marked in a red color.

In addition to the approaches mentioned for controlling FSG there is also a possibility to exclude the COSY data from the process of FSG as a user option (see Fig. 2.10, the group **Use Connectivities from...**). In some cases, especially those when the COSY data contain many NSCs and at the same time the HMBC data are rich enough, the exclusion of the COSY data accelerates the solution of the problem. Removal of weak peaks from COSY and HMBC spectra and elongation of all COSY connectivities up to three bonds (correlations of ${}^4J_{\text{HH}}$ type) can also be helpful. The presence of NSCs in the COSY data can sometimes be detected by repeated MCD checking—with COSY data switched *on* and *off*.

2.3.4 The Strategy of Applying Fuzzy Structure Generation

The possibility of employing several different modes of FSG proves to be a very flexible analytical tool. However, the diversity of modes available is also a source of complexity since the user has to choose the optimal mode when solving a specific problem. Before starting the calculations it is unclear which mode will lead to a solution in a reasonable time.

An attempt was made to answer the question of whether there is a general strategy of structure elucidation using FSG that works best. A set of more than 100 problems were selected where either the HMBC or COSY spectra, or both, contained a total of 1–18 nonstandard connectivities corresponding to a range of coupling constants ${}^nJ_{\text{HH}, \text{CH}}$ where $n = 4–6$. The structures under investigation were all natural products and the number of skeletal atoms in the molecules varied between 15 and 75 skeletal atoms.

For each problem the NMR spectral data were entered into the program and graphically represented as MCDs. The procedure for checking the 2D NMR data for contradictions was then applied to every problem. If the presence of NSCs was

revealed then the program displayed the minimum number of nonstandard connectivities and made an attempt to automatically resolve the contradictions as described above. In successful cases the updated MCDs were displayed with modified connectivities marked by violet color.

As a result of these studies all problems were classified into three sets as follows:

- (1) 53 problems were identified where NSCs were detected and the initial MCDs were updated;
- (2) 34 problems were identified where the program revealed the presence of NSCs but failed to update the MCDs;
- (3) 13 problems were identified where the program failed to detect NSCs.

This classification describes all conceivable results of checking the MCDs. Depending on the results of checking the MCD, various modes or combinations of modes can lead to a solution of the problem. Attempts to solve each problem were made using different FSG modes to investigate possible approaches. The problems for which valid solutions could not be found during the first attempt were eventually solved after utilizing different fuzzy generation options. Logical data pre-processing frequently allowed significant reduction of the number of connectivity combinations to be tested during FSG. Figure 2.30 shows the ratio $\rho = n_{\text{real}}/n_{\text{math}}$, where n_{real} is the number of tested connectivity combinations, $n_{\text{math}} = C_{N_0}^m$ is the theoretically calculated number of combinations. Figure 2.31 examines these combinations in greater detail.

The figures demonstrate that the theoretical number of combinations can be hundreds of billions but the real numbers reduce down to manageable dimensions. For instance, in 20 problems the theoretical number dropped by 10^4 – 10^6 times but the real numbers of combinations still remained rather large. Nevertheless, the speed of the structure generator algorithm was fast enough to solve almost all problems.

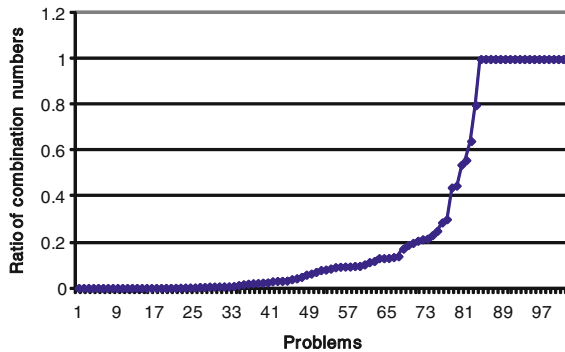
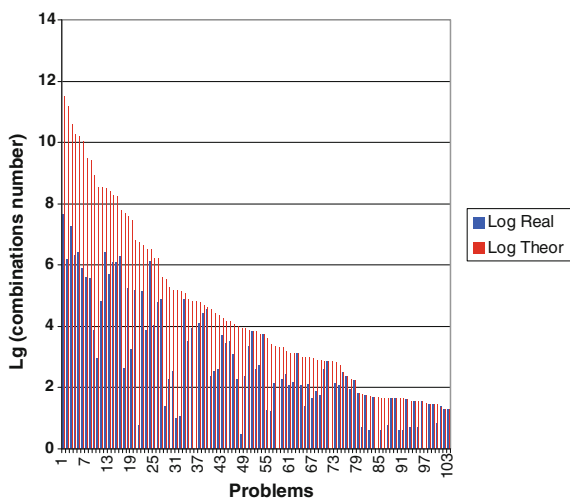


Fig. 2.30 The ratio of the numbers of real connectivity combinations to the numbers of theoretically possible combinations for the problems solved using FSG. The program failed to reduce the number of combinations mainly in those cases when nonstandard connectivities were not detected during checking of the MCD

Fig. 2.31 A plot of the logarithms of the theoretical (red) and real (blue) numbers of connectivity combinations



Fuzzy Structure Generation did, however, fail for the elucidation of structure **d** ($\text{C}_{32}\text{H}_{50}\text{NO}_2$) in Fig. 2.22. The 2D NMR data contain 18 nonstandard connectivities (12 HMBC and 6 COSY nonstandard connectivities; 5 connectivities are of type 5J). The theoretical number n_{math} of connectivity combinations is equal to $\sim 43 \times 10^{12}$ for this case. The difficulty could be circumvented by using the Fragment Mode, but no large appropriate fragment was found in the database during the ^{13}C NMR search. The application of a large UF led to an extremely large set of MCDs with each containing the UF with different distributions of the carbon chemical shifts. As a result these two combinatorial “explosions” hampered problem solving. The solution of such computationally difficult problems will hopefully be eased by further development of the algorithm providing fragment “implementation” in MCDs.

As a result of the studies described, general traits were identified that could help to find appropriate ways to solve a problem. These strategies, as applied to the three problem subsets mentioned above, are described in the following subsections.

2.3.4.1 NSCs Were Identified and the MCD Was Updated

Assuming that the MCD updating process was performed successfully (with the lengths of all NSCs increased) then *strict* structure generation is performed. If an acceptable solution is obtained then it should be checked for stability. FSG with the options $\{m = m_{\min} \div 20; \text{stop at } m = m_g, a = 16\}$ is started from the initial MCD, not the updated MCD. The previously found solution will be confirmed if the first ranked structures for both strict and fuzzy solutions coincide. When an inequality $d_A^{\text{st}}(1) > d_A^{\text{fuz}}(1, m_g)$ is observed ($d_A^{\text{st}}(1)$ —the deviation calculated for the first ranked structure of the solution found by strict structure generation, $d_A^{\text{fuz}}(1, m_g)$ —

the same found by FSG at $m = m_g$), then it is concluded that not all NSCs were lengthened during updating of the MCD and Fuzzy Structure Generation should be repeated with $m_g + 1$ and so on until the minimum value of $d_A^{\text{fuz}}(1, m_g + v)$ and a valid solution is achieved at $m = m_g + v$. The corresponding structure is then considered as the most probable.

An unacceptable solution can be obtained as a result of strict structure generation from the updated MCD, i.e., a solution will be found for either $d_A^{\text{st}}(1) > D_A$, where D_A is a threshold value or an empty structural file is obtained ($k = 0$). In both cases the program is automatically switched to the mode where $\{m = m_{\min} \div 20, \text{ stop at } m = m_g, a = 16\}$. Depending on the m_g values and the complexity of the problem (the size of n_{real} and the calculation time) evaluated during the first stages of solving the problem, the user can initiate FSG with the options $\{m \leq m_0, a = 16\}$, $m_0 = 5, 10$ or 15 to obtain the most reliable solution.

2.3.4.2 NSCs Were Identified but the MCD Failed to Be Updated

If the program identified NSCs but failed to update the MCD, then FSG is one manner by which to solve such a problem. Since the program only displays the minimum number of NSCs while their associated lengths remain unknown, FSG with the options $\{m = m_{\min} - 20, \text{ stop at } m = m_g, a = 16\}$ should be used. The real numbers of the connectivity combinations, n_{real} , are displayed, as well as the number of combinations for a given $m = m_g$, and the approximately predicted time for structure generation allows the user to easily evaluate the complexity of the problem and the suggested time for execution. If *Mode 4* can be applied based on acceptable time estimates then it should be used.

2.3.4.3 NSCs Were Not Detected

If nonstandard connectivities were not revealed by checking the MCDs then there are two ways to interpret this result: either the 2D NMR data is free of nonstandard connectivities or the implicit NSCs are present but the program failed to detect them. Both of these situations are covered by FSG with the options $\{m = 0 \div 20, \text{ stop at } m = m_g, a = 16\}$. If NSCs are indeed absent from the 2D NMR data then structure generation is performed with $m = 0$ with a nonzero output file and the deviation values allow the user to determine whether the solution determined is acceptable. Obtaining deviation values that exceed the threshold for d_A , or deriving an empty output file after spectral filtering, both serve as hints to the presence of latent nonstandard connectivities.

When NSCs are not detected by logical data analysis then the number of connectivity combinations that must be tested during FSG cannot be reduced and it is equal to $C_{N_0}^m$, $m = 1, 2, 3, \dots$ at each m th step of the FSG process. This situation can cause significant difficulties due to an unmanageable number of connectivity

combinations needing to be processed; as discussed previously, both FF and UF can assist in this situation.

It is hardly possible to describe all of the nuances associated with FSG since these depend on each 2D NMR data set associated with a given problem. A series of examples illustrating the strategies leading to valid solutions with the minimum number of user assumptions will be presented in Chap. 5.

2.3.5 Is There an Alternative to Fuzzy Structure Generation?

Some researchers suggested that it was possible to overcome the problem of NSCs by setting default values for $^4J_{\text{CH}, \text{HH}}$ for *all* COSY and HMBC correlations observed in the 2D NMR spectra. It was important to answer the question: To what extent can the lengthening of *all* 2D NMR correlations act as a method for contradiction resolution in 2D NMR data? A study was undertaken to answer this question [7]. In this study an attempt was made to identify, in a quantitative manner, how the structure generation time increases and the amount of structural information obtained decreases if only correlations in the ^{2-4}J range were allowed.

Analysis of the results of this study showed that even in the case of small molecules the output file size increases considerably when the ^{2-4}J couplings are set as default. The generation time increases by many times to hundreds or even tens of thousand times greater. For one of the studied problems the size of the output file increased from 2 to ca. 3,000 structures, while the generation time increased by 6.5 million times! The main conclusion is that the lengthening of *all* correlations should be rejected as a general method of solving problems arising from the presence of nonstandard correlations in 2D NMR data.

References

1. Thongbai B, Surup F, Mohr K, Kuhnert E, Hyde KD, Stadler M (2013) Gymnopalynes A and B, chloropropynyl-isocoumarin antibiotics from cultures of the basidiomycete *Gymnopus sp.* J Nat Prod 76(11):2141–2144. doi:10.1021/np400609f
2. Benie AJ, Sørensen OW (2007) HAT HMBC: a hybrid of H2BC and HMBC overcoming shortcomings of both. J Magn Reson 184(2):315–321
3. Krishnamurthy V, Russell D, Hadden C, Martin GE (2000) 2J, (3)J-HMBC: a new long-range heteronuclear shift correlation technique capable of differentiating (2)J(CH) from (3)J(CH) correlations to protonated carbons. J Magn Reson 146(1):232–239
4. Nyberg NT, Duus JØ, Sørensen OW (2005) Heteronuclear two-bond correlation: suppressing heteronuclear three-bond or higher NMR correlations while enhancing two-bond correlations even for vanishing 2J(CH). J Am Chem Soc 127(17):6154–6155
5. Sprang T, Bigler P (2003) A new technique for differentiating between 2J(C, H) and 3/4J(C, H) connectivities. Magn Reson Chem 41(3):177–182

6. Molodtsov SG, Elyashberg ME, Blinov KA, Williams AJ, Martin GM, Lefebvre B (2004) Structure elucidation from 2D NMR spectra using the StrucEluc expert system: detection and removal of contradictions in the data. *J Chem Inf Comput Sci* 44:1737–1751
7. Elyashberg ME, Blinov KA, Molodtsov SG, Williams AJ, Martin GE (2007) Fuzzy structure generation: a new efficient tool for computer-aided structure elucidation (CASE). *J Chem Inf Model* 47(3):1053–1066
8. Elyashberg ME, Blinov KA, Williams AJ (2009) A systematic approach for the generation and verification of structural hypotheses. *Magn Reson Chem* 47(5):371–389. doi:[10.1002/mrc.2397](https://doi.org/10.1002/mrc.2397)
9. Blinov KA, Elyashberg ME, Martirosian ER, Molodtsov SG, Williams AJ, Sharaf MMH, Schiff PLJ, Crouch RC, Martin GE, Hadden CE, Guido JE, Mills KA (2003) Quindolinocryptotackeine: the elucidation of a novel indoloquinoline alkaloid structure through the use of computer-assisted structure elucidation and 2D NMR. *Magn Reson Chem* 41:577–584
10. Martin GE, Hadden BD, Russell CE, Kaluzny DJ, Guido JE, Duholke WK, Stiemsma BA, Thamann TJ, Crouch RC, Blinov KA, Elyashberg ME, Martirosian ER, Molodtsov SG, Williams AJ, Schiff PLJ (2002) Identification of degradants of a complex alkaloid using NMR cryoprobe technology and ACD/structure Elucidator. *J Heterocycl Chem* 39:1241–1250
11. Elyashberg ME, Blinov KA, Molodtsov SG, Williams AJ (2013) Structure revision of asperjinone using computer-assisted structure elucidation methods. *J Nat Prod* 76:113–116
12. Elyashberg ME, Blinov KA, Molodtsov SG, Williams AJ (2012) Elucidating “undecipherable” chemical structures using computer assisted structure elucidation approaches. *Magn Reson Chem* 50:22–27
13. Elyashberg ME, Williams AJ, Martin GE (2008) Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. *Prog Nucl Magn Reson Spectrosc* 53(1/2): 1–104
14. Elyashberg ME, Williams AJ, Blinov KA (2012) Contemporary computer-assisted approaches to molecular structure elucidation, vol 1. *New Developments in NMR*. RSC Publishing, Cambridge
15. Berger S, Braun S (2004) 200 and more NMR experiments: a practical course. Wiley, New York
16. Collins DO, Reynolds WF, Reese PB (2004) New cembranes from *Cleome spinosa*. *J Nat Prod* 67:179–183
17. Mensah AY, Houghton PJ, Bloomfield S, Vlietinck A, Berghe DV (2000) Known and novel terpenes from *buddleja globosa* displaying selective antifungal activity against dermatophytes. *J Nat Prod* 63:1210–1213
18. Wellington KD, Cambie RC, Rutledge PS, Bergquist PR (2000) Chemistry of sponges. 19. Novel bioactive metabolites from *Hamigera tarangaensis*. *J Nat Prod* 63:79–85
19. Oliveira JHHL, Grube A, Köck M, Berlinck RGS, Macedo ML, Ferreira AG, Hajdu E (2004) Ingenamine G and cyclostelletamines G-I, K, and L from the new Brazilian species of marine sponge *Pachychalina sp.* *J Nat Prod* 67:1685–1689

<http://www.springer.com/978-3-662-46401-4>

Computer-Based Structure Elucidation from Spectral
Data

The Art of Solving Problems

Elyashberg, M.E.; Williams, A.J.

2015, XVI, 447 p. 536 illus., 352 illus. in color.,

Hardcover

ISBN: 978-3-662-46401-4