Computer assisted structure elucidation of natural products using two-dimensional NMR spectroscopy†

Marcel Jaspars

Marine Natural Products Laboratory, Department of Chemistry, University of Aberdeen, Meston Walk, Old Aberdeen, UK AB24 3UE

Received (in Cambridge) 28th September 1998

1 Introduction

The ideal of computer assisted structure elucidation (CASE) is to generate, exhaustively and without redundancy, all possible structures that are consistent with a particular set of spectroscopic data. The aim is to achieve this goal with the minimum amount of human intervention. In natural product drug discovery the major bottleneck has always been structure elucidation, and it is with this in mind that the development of CASE began. Great advances have been made in this field in the past few years and programs that achieve most of these goals are currently available. The scope of this review is primarily to discuss the role of 2D NMR spectroscopy in CASE. Since this is a review for natural product chemists, the focus will be mainly on the use of these programs in solving complex natural product structures. The intended purpose of this review is not to be fully comprehensive, but rather to highlight different approaches to CASE and significant recent developments. It is also not the intention of the author to focus on the mathematical aspects of CASE, but instead to explain the concepts used in basic terms. The first part of the review highlights alternative approaches which either do not make use of 2D NMR data, or make use of them in a limited way. This is followed by the methods a natural product chemist may apply to elucidate the structure of a complex natural product. The next part of this review is structured in the same way that a CASE program fits together, and begins with a general description of the system, followed by details of the component parts.

† Abbreviations used: 2D NMR, Two Dimensional NMR; ACF, Atom Centred Fragment; CASE, Computer Assisted Structure Elucidation; CD, Circular Dichroism; COCOA, Structure Generator for SESAMI; COSY, Correlated Spectroscopy; CPU, Central Processing Unit; DEPT, Distortionless Enchancement by Polarisation Transfer; DQF-COSY, Double Quantum Filtered COSY; GENOA, Generation with Overlapping Atoms; HETCOR, Heteronuclear Correlation Spectroscopy; HMBC, Heteronuclear Multiple Bond Correlation; HMQC, Heteronuclear Multiple Quantum Coherence; HMQC-TOCSY, Heteronuclear Multiple Quantum Coherence-Total Correlation Spectroscopy; HSQC, Heteronuclear Single Quantum Coherence; INADEQUATE, Incredible Natural Abundance Double Quantum Transfer Experiment; MOLGEN, Molecular Generator; NOESY, Nuclear Overhauser Enhancement Spectroscopy; ORD, Optical Rotatory Dispersion; PRUNE, Algorithm in SESAMI that reduces the set of ACFs; TOCSY, Total Correlation Spectroscopy.

2 Alternative approaches

Other approaches to CASE, without resorting to 2D NMR data, have been tried using ^{13}C NMR data alone. One example of this type of system is Richert's SpecSolv[1] which is a new module of the NMR database SpecInfo.[2] SpecInfo has used data from thousands of compounds to calculate typical chemical shifts for a carbon with a particular set of neighbours (a substructure). SpecSolv allows the user to enter the ^{13}C NMR spectrum of the unknown, without having to give the molecular formula, and structures matching these chemical shifts are returned. For 80% of all compounds containing only C, H, N, O, S, P and halogens, the correct structure is derived. The program relies on a subspectrum search, which is then translated to a collection of substructures. The substructures are assembled to give the greatest degree of overlap, and the ^{13}C chemical shift is calculated for each generated structure. The structure which gives the correct ^{13}C NMR spectrum is returned to the user as the most likely candidate structure. With 'exotic unknowns' such as complex natural products no final structure can be proposed by SpecSolv due to the lack of subspectral matches. In addition, two carbons with the same neighbouring groups, but in different conformations may have very different chemical shifts, and this may confound the subspectrum search. Although ^{13}C shift based programs are likely to find great utility in a synthetic laboratory with a high turnover of compounds, it is unlikely to fulfil the ideal of CASE stated above.

Another approach to CASE, which does incorporate the use of 2D NMR data, is to compose a new NMR pulse sequence which enables the direct determination of proton spin systems in a molecule.[3] The generation of proton spin systems is also possible using a graph theoretical method which determines a C–C connectivity matrix for protonated carbons by the direct determination of the matrix product of ^{1}H–^{1}H COSY and ^{1}H–^{13}C COSY (1 bond) spectra.[4] These last two methods are useful in generating spin systems only, but they do not allow the generation of complete structures without the use of further long range data. This review will therefore focus on CASE programs which are able to use routinely available 2D NMR data (*e.g.* ^{1}H–^{1}H COSY, HMQC or HSQC, HMBC, NOESY and INADEQUATE).

3 The human thought process

In order to understand how CASE programs are constructed, it is important to appreciate how a spectroscopist will elucidate a structure from spectroscopic data. The process is summarised in Fig. 1.[5] Normally the molecular formula is derived from a combination of ^{13}C NMR, DEPT and MS data. Using IR, UV and ^{13}C NMR the functional groups can be proposed, and ^{1}H NMR coupling data or 2D NMR correlations are used to assemble substructures. These are then combined into 'working structures' which are all possible combinations of the substructures. These are then checked for consistency with the 2D NMR data and MS fragmentations *etc*. The ^{13}C chemical shifts of the surviving structure(s) are then compared with literature, database or predicted values to confirm the 2D structure of the molecule. To determine the relative stereochemistry of the molecule, ^{1}H coupling constant (J) and Nuclear Overhauser

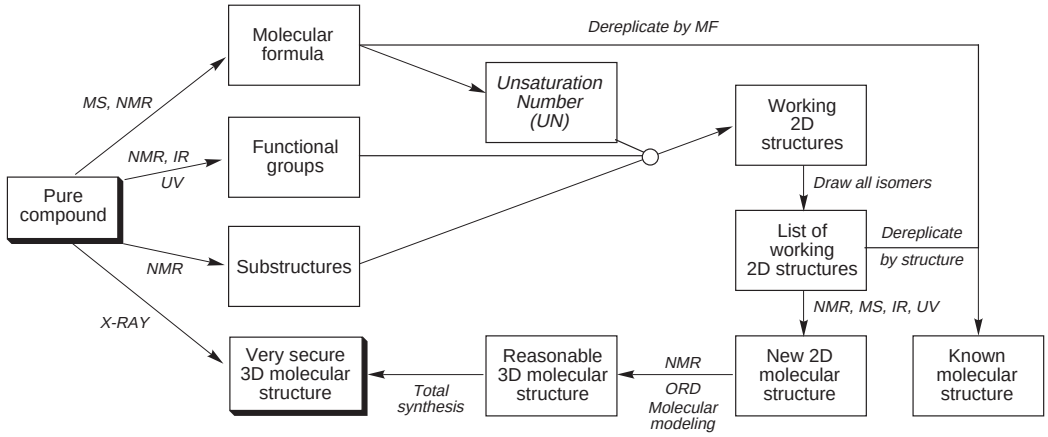

Fig. 1 Strategy for structure elucidation from spectroscopic data. (From 'Organic Structure Analysis' Crews, Rodriguez and Jaspars, used with permission from Oxford University Press.)

Enhancement (NOE) data are used. The absolute stereochemistry can then be determined by a variety of methods such as ORD/CD, derivatisation or degradation.[6] As early as possible in this process, the chemist needs to determine whether the unknown in question has previously been described, a process known as dereplication. This can be performed using a combination of molecular formula, substructures and chemical/structural databases.[7] Once it has been established that the compound in question has not been reported before, the process of structure elucidation as depicted in Fig. 1 can begin. The problem with this approach is that it depends on assumptions based on past experience (heuristics). The experience and prejudices of the spectroscopist will guide the solution of an unknown towards a particular structural type, for instance using biosynthetic rules. There may be other solutions which are consistent with the spectroscopic data, but these may be excluded for various reasons by the spectroscopist. One major aim of a CASE program is to make certain that all structures consistent with a given data set are generated without prejudice. In order to do so efficiently it may be necessary to steer away from heavy reliance on chemical shift data, and instead use connectivities obtained from 2D NMR spectra. It will become obvious that the most successful CASE programs tend to follow the human thought process closely.

Two common strategies are employed when elucidating organic structures, one involving direct C–C correlations from a 2D INADEQUATE spectrum, the other using C–C connectivities inferred from C–H and H–H data. The INADEQUATE strategy is summarised in Fig. 2. The process is started by obtaining the ^{13}C NMR spectrum and multiplicities, followed by the acquisition of the 2D INADEQUATE spectrum. This allows the construction of the carbon skeleton of the molecule relatively easily, except in the case of extreme overlap in the ^{13}C spectrum. If heteroatoms are present in the skeleton of the molecule, these can be inferred from ^{13}C chemical shifts, and connectivity between the carbons bridged by a heteroatom is deduced by obtaining direct C–H correlations to assign ^{1}H resonances followed by the use of long range (2–3 bond) C–H correlations. Relative and absolute stereochemistry are then obtained as shown in Fig. 1. The main problem of this approach is the inherent insensitivity of the INADEQUATE experiment, which dictates that a large amount of sample is needed and that it must be soluble in a small amount of solvent. In the case of menthol, a 6 molar solution in $CDCl_3$ is needed to obtain a 2D INADEQUATE spectrum in 24 h. This is clearly not practicable in all cases, but recent developments mean that this experiment can now be performed using as little as 11 mg of a compound of molecular formula $C_{21}H_{32}O_3$ in a reasonable time (62 h).[8] This was achieved using a low volume NMR probe (40 μL) at 500 MHz and special analysis software which enabled the INADEQUATE correlations to be picked from the noise, giving a roughly 10 fold improvement in sensitivity.[9] Even if the necessary equipment and software is routinely available, problems still exist with this method. The time required is still greater than alternative strategies, and a ^{1}H–^{13}C one-bond correlation spectrum (*e.g.* HMQC or HSQC) must still be acquired to assign the proton chemical shifts. In addition, if the skeleton is broken by heteroatoms, then a long range ^{1}H–^{13}C NMR or NOESY spectrum must also be obtained to complete the structure. Another problem occurs when the $^{1}J_{CC}$s vary widely within the molecule, in which case more than one INADEQUATE spectrum needs to be obtained. In brief, this strategy will be used in special cases only, for instance, when there is extreme overlap in the ^{1}H NMR spectrum and good resolution in the ^{13}C NMR spectrum.

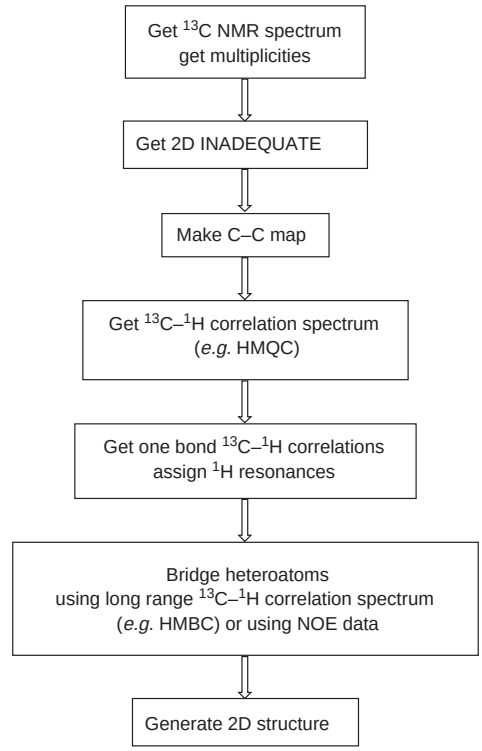

Fig. 2 A possible structure elucidation strategy using C–C correlation data.

The alternative strategy involves the use of more 2D NMR experiments, but these can be obtained in a reasonable time using inverse detected techniques on a multimilligram sample, and this strategy is outlined in Fig. 3. The process is started by obtaining ^{1}H and ^{13}C NMR data with multiplicities and

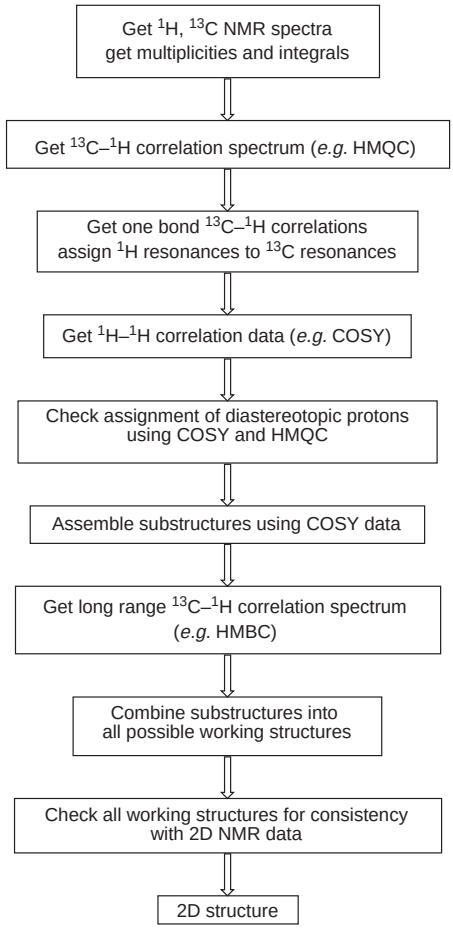

Fig. 3 A possible structure elucidation strategy using H–H and H–C correlation data.

integrals, after which a one bond ^{1}H–^{13}C NMR spectrum (HMQC, HSQC) is obtained allowing the assignment of protons to particular carbons. In cases of severe spectral congestion in the ^{1}H NMR spectrum, an HMQC–TOCSY spectrum can be used to aid C–H assignment.[10,11] ^{1}H–^{1}H correlation data (*i.e.* COSY, TOCSY) are obtained next for two reasons, they allow confirmation of diastereotopic protons inferred from the HMQC or HSQC, and spin systems can be constructed. These substructures can then be combined with the assistance of a ^{1}H–^{13}C NMR long range NMR spectrum (*i.e.* HMBC). At this stage all possible working structures must be constructed and their consistency checked with all the 2D NMR data. Again, relative and absolute stereochemistry are then obtained as shown in Fig. 1. There are problems with this approach, and most of these arise from the ambiguity of the data. In many cases in the ^{1}H–^{1}H COSY, it is possible to distinguish between two, three and four bond correlations by the strength of the cross peaks, but in some cases, four bond are stronger than three bond correlations, thus confusing the issue. In the HMBC spectrum it is impossible to distinguish between two and three bond C–H correlations. In addition to this, the angular dependence of $^{3}J_{HH}$ and $^{3}J_{CH}$ means that some expected cross peaks will not appear in the ^{1}H–^{1}H COSY and HMBC NMR spectra respectively. The resolution in inverse detected NMR spectra is poorer in the ^{13}C dimension than for HETCOR spectra, increasing the likelihood of ambiguous assignments. The resolution in the ^{13}C NMR dimension can be improved using linear prediction, thus alleviating this problem somewhat. A major advantage of using inverse detected experiments is the eight-fold improvement in sensitivity of HMQC/HSQC over HETCOR.[12] Using modern high field NMR instruments, the entire dataset (^{1}H, ^{13}C, DEPT, HMQC or HSQC, COSY, HMBC, NOESY) necessary to elucidate the structure of an unknown natural product of reasonable molecular weight (< 1 kDa) can be obtained in one overnight run on a 10–100 mM sample. In fact, the routine availability of pulsed field gradients has removed the need for phase cycling in 2D NMR experiments, and if good signal to noise can be obtained for a one transient ^{1}H NMR spectrum, single transient 2D NMR spectra can be acquired in very short times (< 1 h). This then means that the time limiting factors are the ^{13}C observed 1D spectra. This strategy using H–H and C–H correlation data is most routinely employed by spectroscopists as it is the most time efficient method. Thus the most successful CASE programs must use these experiments, and must be able to deal with the inherent ambiguities of the data.

4 Structure of a CASE program

A CASE system is composed of several parts (Fig. 4) which will all be discussed in detail below. The first part is the input of the

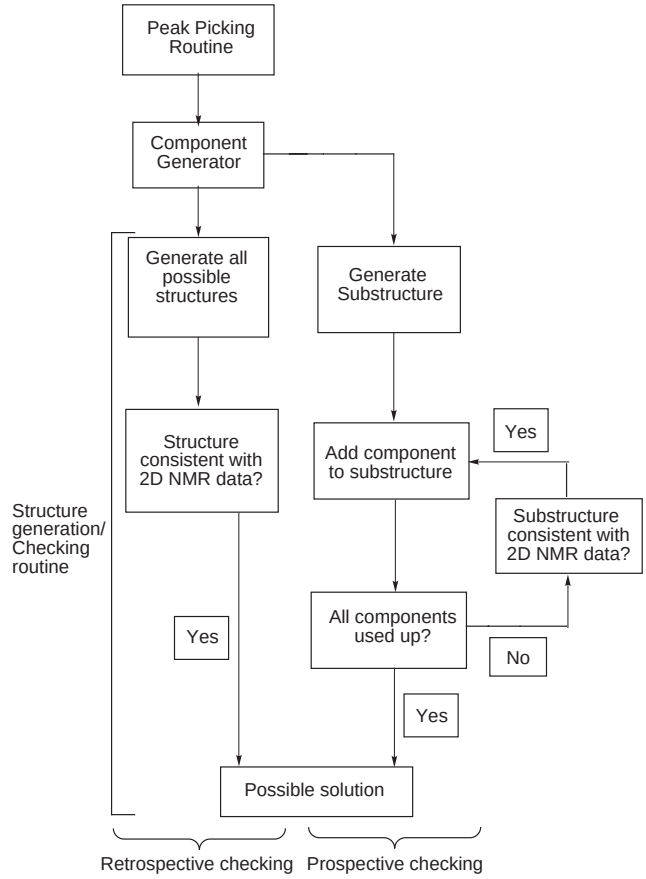

Fig. 4 Components of a CASE program.

spectra, or 'peak-picking'. This can rely on the skill of a spectroscopist who translates the cross peaks of a 2D spectrum into correlations, or ideally on a sophisticated peak picking program. Next, components can be generated (*e.g.* CH_3, CH_2–O *etc.*), and these are fed into the most important part of the program, the structure generator. This will then exhaustively generate all possible structures from these components, and it is this part of the CASE program that will take the greatest amount of CPU time, and this is also where the greatest time savings can be made by the use of efficient algorithms. The generated structures are checked for consistency with the 2D NMR data. Generated substructures can be checked during the process of structure generation (prospective checking) or after all complete structures have been generated (retrospective checking). Clearly, prospective checking will be faster, as those sub-structures that are not consistent with the 2D NMR data are removed from the structure generation process. In the case of

retrospective checking, a combinatorial explosion occurs for exhaustive structure generation, even for molecules of a moderate size. Another very important aspect of a structure generator is checking for redundancy, meaning that it checks that it is not producing the same structure more than once. At the end of all this the program should return all those structures that are consistent with all the spectroscopic data entered.

5 Components of a CASE program

5.1 Peak picking routines

Once spectra of an unknown compound have been acquired, the first task is their analysis to convert the data into a more computer digestible form. Most current CASE programs (*e.g.* CHEMICS,[13] GENOA,[14] SESAMI[15] and LSD[16]) rely on the spectroscopist determining correlations from the cross peaks of a 2D spectrum, either by hand, or using available software. Much of the peak picking software has been designed for analysis of 2D, 3D and 4D NMR spectra of biomolecules,[17,18] and is far in excess of that needed for natural product structure elucidation. With current inverse detected NMR experiments and the use of pulsed field gradients and linear prediction it is now possible to obtain 2D NMR spectra with high resolution in both dimensions, good signal to noise and low artefact levels in a reasonable amount of time on a small sample. This in turn makes the task of the peak picking program a great deal simpler, as the chances of a noise or artefact peak being picked are considerably reduced. One method is to search the entire 2D spectrum for peaks with the correct peak shape to distinguish them from noise. The problem with this is that the search routine will encounter nothing for the greater part of its time, and this is therefore very time inefficient. An alternative approach is to search only those regions of the spectrum where peaks are expected (*i.e.* at the crossing points of ^{1}H resonances with either ^{1}H or ^{13}C resonances), and this has been achieved in different ways for homonuclear (^{1}H–^{1}H) and heteronuclear (^{1}H–^{13}C) spectra. In the case of a ^{1}H–^{1}H COSY spectrum, a 45° projection of a ^{1}H 2D homonuclear *J* resolved spectrum is obtained.[19] This is effectively a proton decoupled proton spectrum, and gives a single chemical shift for each magnetically non-equivalent proton in the molecule. These chemical shift values are used to construct a 'shift grid' which is used by the algorithm to search the intersections of the grid. The program also tests for mirror symmetry to remove noise peaks, and also gives a confidence value for the correlations found. A similar approach was taken in the program CISOC–SES (Computerised Information System for Organic Chemistry–Structure Elucidation System).[20] Here the program obtains the ^{13}C NMR chemical shifts together with multiplicities, and the search routine then uses these chemical shift values to investigate regions of the HMQC or HSQC spectrum for the correct number of attached protons, *i.e.* one peak for a CH, two for a CH_2 and one for a CH_3. This approach is subject to interference by noise, but the program is interactive and allows the operator to remove suspicious peaks. Later in the routine, diastereotopic pairs determined this way are cross checked for large geminal correlations in the ^{1}H–^{1}H COSY spectrum. The list of ^{1}H chemical shift values is then used to construct a 'shift grid' in the analysis of the other homonuclear and heteronuclear NMR spectra. Both approaches work well in practice and greatly reduce the time taken to analyse complex spectra. A major difficulty with both manual and computer peak picking is when spectral overlap occurs. When a given cross peak in a 2D NMR spectrum can be related to more than one proton or carbon chemical shift, ambiguity arises. This problem is best solved by designing the structure generator so that it will accept ambiguous assignments, as will be seen below. A similar problem occurs when there is symmetry in the molecule, so that a given chemical shift relates to more than one carbon atom. In some cases this can be resolved through the use of ^{1}H NMR integrals, but in many cases operator input is required.

5.2 Generation of components

After analysis of the spectra has been accomplished by a peak picking routine, the next step is to produce a list of possible components present in the molecule. Such a list must contain all possible components, without exception, that are consistent with the molecular formula and spectral data. This is the way in which CHEMICS and SESAMI (Systematic Elucidation of Structure Applying Machine Intelligence) operate. This list of components will often contain more invalid components (*i.e.* those not present in the unknown) than valid ones. In CISOC–SES and LSD the sum of all the components is the molecular formula (*i.e.* each atom from the molecular formula is included once only). Clearly the latter approach is more productive and will be described first. In these cases components contain information about a particular carbon atom, such as hybridisation state, number of attached protons, and in some cases attached heteroatoms. The simplest way to define these components is for a skilled spectroscopist to assess the ^{13}C NMR and multiplicity data and construct a list containing atom number, element, chemical shift, hybridisation state and number of attached hydrogens. This is the approach taken by the program LSD (Logic for Structure Determination),[16] as it simplifies the task for the program and reduces the possible structures that can be generated. This has been done because it is difficult to write a computer algorithm that can always determine the hybridisation state of a carbon atom. The problem is particularly acute in the ^{13}C NMR spectral region close to 100 ppm where we may get shielded sp^2 carbons or sp^3 acetalic carbons. A way round this has been found by the writers of the CISOC–SES program.[21] Their approach is to ignore hybridisation, and instead use the concept of 'free bonds'. An example would be that information from ^{13}C and DEPT NMR determines that a particular carbon has only one attached proton, and it therefore has three unsaturated valences or 'free bonds'. These free bonds can be used to form one triple bond, one double and one single bond or three single bonds. This approach is very versatile and removes the pitfalls associated with the determination of the hybridisation state of a particular carbon. A minimal knowledge of chemical shifts is included to enable the construction of units such as carbonyl groups. User intervention is possible here to add or alter any of these components. When there is a degree of symmetry in the molecule (*i.e.* more carbons in the molecular formula than ^{13}C NMR chemical shifts), the program can use ^{1}H integrals to ascertain this, but often the best solution is for the operator to enter this information explicitly. A CASE program that has been designed to cope with symmetry is SESAMI.[15] This program uses atom centred fragments (ACF) as its components for structure generation. An ACF contains information about the central element, its attached hydrogens, and each of the bonds to the central atom, *e.g.* =CH–CH_2–O–, and in total about 5100 ACFs are recognised. SESAMI uses the molecular formula, and ^{1}H and ^{13}C NMR spectra to generate an exhaustive list of ACFs which are then reduced by a routine called PRUNE based on a series of rules. A spectroscopist can then inspect the shortlist of ACFs and add or delete some of these based on experience.

The most complex approach of all is taken by the program CHEMICS, mainly for historical reasons. CHEMICS is one of the oldest CASE systems and relied initially on IR, ^{1}H and ^{13}C NMR data, and later grew to include 2D NMR data.[13,22] In CHEMICS a list of 86 secondary components [*e.g.* $(CH_3)_2C$=] is expanded into a list of 630 tertiary components which describe allowed bonding partners of the secondary components.[23] These components are chosen based on spectral data and a complex, iterative set of rules. The dangers of systems which employ larger components as starting points for structure

generation are that some structures might be missed because of the absence of a component from the list of allowed components. One advantage of these systems (SESAMI, CHEMICS) is that they have easy mechanisms by which known components/substructures can be excluded from the structure generation process, or can be forced to be included. The inclusion and exclusion of substructures like this is also possible for CISOC–SES and LSD, by the operator providing a connectivity list. In all the above cases the danger of operator prejudice may become apparent, as a spectroscopist may decide to include or exclude a particular substructure from the structure generator, and thus miss generating the correct solution. A final point to be made here is that on many occasions, the molecular formula is not easily determined early on in the structure elucidation process. In those cases, a spectroscopist may decide to carry on without this information, and is often able to propose a viable solution. Some CASE programs can construct a list of components suitable for structure generation without using any further information.[15,21] The ability to perform CASE without a molecular formula is being developed and is present in a recent implementation of CISOC–SES.[24] This will be useful to aid the spectroscopist in the early part of a structure elucidation, where only limited information is available, and may help in the construction of substructures.

5.3 Structure generation/consistency checking

The purpose of a structure generator is to use the components generated by one of the methods mentioned above and produce an exhaustive list of all possible structures without redundancy, and without missing out any plausible structures. Several structure generators have been developed which accept components or substructures, and together with other input from the spectroscopist will generate an exhaustive list of structures without redundancy; one example of a very good structure generator is MOLGEN.[25] Using this method is very time consuming, as all possible structures are generated, to be checked by the spectroscopist for consistency with the 2D NMR data (*c.f.* retrospective checking in Fig. 4). The method also suffers because without a more selective structure generation strategy, a combinatorial explosion occurs, taking up prohibitive amounts of CPU time. A more productive strategy is to check the consistency of substructures with 2D NMR data as the structure is being generated (Fig. 4, prospective checking). The field of structure generation is enormous (*e.g.* the work of Bangov *et al.*[26,27]), and very dependent on graph theory,[28] and there will not be space to do it justice here. This review will concentrate on those methods that are part of successful CASE systems. For SESAMI and CHEMICS, the shortlist of components contains many components that are not present in the unknown sample. The first task of the structure generator is therefore to select groups of components from this list whose sum totals the molecular formula of the unknown. COCOA, the structure generator of SESAMI, utilises this set of shortlists together with constraint information obtained from 2D NMR data and user defined substructures.[15] CHEMICS creates its set of component shortlists and reviews and updates these throughout the structure generation process to generate complete candidate structures.[13] This program is able to utilise data from 2D INADEQUATE, HMQC or HSQC and ^{1}H–^{1}H COSY experiments, and is currently being updated to enable it to use ambiguous HMBC data, but this function has not yet been described.[13] The realisation that a combination of HMQC or HSQC, ^{1}H–^{1}H COSY and HMBC data could be used to generate in effect pseudo 'C–C' bond data was made early on (Fig. 5).[15,16,21] The use of $^{3,4}J_{HH}$ from the ^{1}H–^{1}H COSY or the use of $^{2,3}J_{CH}$ from the HMBC together with the $^1J_{CH}$ from the HMQC or HSQC yields C-C and C–C–C information in each case. This information is ambiguous, though in most, but not all, cases ^{1}H–^{1}H COSY correlations corresponding to three and

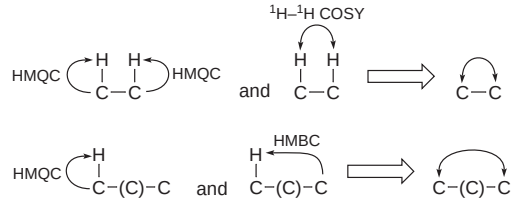

Fig. 5 Carbon–carbon correlation deduced from 2D NMR spectra.

four bond can be distinguished from each other. In the HMBC the situation is far worse, because in most cases no distinction can be made between two and three bond C–H correlations. In the cases where a three bond C–H correlation can be inferred, *e.g.* for a proton correlated to a methyl carbon on a quaternary carbon (CH_3–C–CH), this can be coded explicitly. Despite these difficulties, the use of a sufficient number of these topological distance constraints derived from HMBC data can be used to generate a single unique structure. The structure generator of SESAMI is able to take these ambiguities into account, by recognition of overlapping substructures, and also remove any atom centred fragments that are no longer consistent with the substructure produced. Examples of the use of SESAMI are given, but the structure generation algorithm is not described in detail.[15]

LSD and CISOC–SES incorporate the use of ambiguous data extremely well, whether the ambiguity arises from the nature of the 2D NMR experiments or from the assignment of 2D NMR cross peaks to more than one unique pair of chemical shifts. Both programs also use data prospectively to narrow down the search, and make use of weighting functions to achieve the correct solution in the minimum amount of time. Other problems that must be dealt with are the rapid recognition of duplicated substructures during the structure generation process and the removal of chemically impossible structures. A routine for the removal of some chemically impossible structures from the generation process is included in LSD.[29] LSD accomplishes structure generation by treating it as a constraint satisfaction problem which is solved using constraint propagation, intelligent backtracking and ordered searching techniques.[30] The strategy used is to assign atom values from the molecular formula to atoms of the substructure (node variables) until an assignment takes place that is not consistent with all the previous assignments, at which stage backtracking takes place, and another atom value is assigned to the preceding variable. What this means in plain terms is that a substructure is constructed, at which stage a new component is attached (see Fig. 4), and checked for consistency with all the 2D NMR data. All available components are tried for consistency and if none are found, the previous substructure is rejected or altered, and the process of adding components is continued. In this way all possible structures consistent with the 2D NMR data are composed. In addition to this backtracking, once a consistent substructure is generated, the effect that this has on other constraints is determined and recorded for further use, a process called constraint propagation. Apart from 2D NMR correlations, other, more basic constraints are also used by LSD. One example is an sp^2 carbon whose bonding partners are all sp^3 carbons; the solution is rejected as no double bond can be formed. A second example is the determination of the total number of rings in the combined substructures. If this number exceeds the maximum number of rings possible in the final structure then the set of substructures is rejected. Using the algorithms described will certainly lead to all possible structures consistent with the constraints, but there may be duplication and also the correct solution may be generated last. Both these problems can be solved by the use of weighting factors and ordering the structure generation process, which increases the likelihood of the correct solution being found first and also reduces the number of redundant structures. The first problem is

where to start, and possibilities include the use of the component participating in the highest number of constraints or a random first choice followed by adding the component with the largest number of constraints. In LSD the authors have chosen to use more than one criterion, and a weighting function is calculated from the domain size of the variables, number of attached constraints and number of free valences amongst others. The component with the highest score is then chosen to begin the process of structure generation. Approximate distances are used to order the structure generation process. If 2D NMR data indicates there is a direct connection between two atoms the distance is set to 1, if there is a long range correlation the value is set to 2, and if there is no correlation the value is set to ∞. The structure generator favours the creation of bonds between components whose approximate distance is 1 or 2, and this way the rapid generation of the most likely solutions is frequently guaranteed. Good examples of the use of LSD are given in the current literature.[31,32]

As mentioned above, CISOC–SES is built around the concept of unsaturated valences or 'free bonds' to avoid the problems in the assignment of hybridisation states. The free bonds of each component are included in the free bond connection matrix which is the basis of the structure generator. Information is included in this matrix similar to the approximate distances in LSD, and in CISOC–SES the free bonds are either allowed to connect or not. The peak picking routine provides connectivity information which is used to reduce the size of the free bond connection matrix. This information is encoded as follows: correlated atom pair; number of intervening bonds between the atoms (*e.g.* two or three for HMBC data); bond type (*i.e.* 1 for single, 2 for double *etc.*, and 0 for HMBC data where bond type is meaningless). Known two or three bond C–H correlations may be entered explicitly, but this is not recommended as this may lead to exclusion of the correct structure, so it is often wisest to leave this parameter at 2–3 bonds. The ^{1}H–^{1}H COSY peaks are treated as in Fig. 5, and are used to construct C–C one bond correlations. If weak correlations are present that may be due to four or five bond H–H correlations, then this can be accounted for either by the program or by the operator. If two protons do not have a ^{1}H–^{1}H COSY correlation between them, then they are forbidden to connect in the subsequent structure generation. This could lead to the exclusion of certain correct structures from the generation process, for instance when the coupling constant between the protons is 0 Hz because the dihedral angle between them is 90°, and for this reason this algorithm can be switched off, with a concomitant increase in structure generation time. Vicinal protons that show no cross peak in the ^{1}H–^{1}H COSY NMR spectrum will show a cross peak in the NOESY spectrum, and this information can be used to avoid excluding connectivities in these cases.[20] One other occasion when a ^{1}H–^{1}H COSY cross peak might be missed is when it is near the diagonal, as this can occur even in a DQF–COSY. In order to avoid this CISOC–SES creates a pseudo-connectivity between protons whose chemical shift difference is below a pre-set value, and this is used as a loose constraint during the structure generation. The free bond matrix is further reduced in size by the use of user defined substructures and C–C one bond connectivities determined from a combination of the ^{1}H–^{1}H COSY and HMQC or HSQC spectra, and these are called fixed connections. This reduction in size, the removal of two free bonds per bond generated reduces the time taken by the structure generator significantly, and several other rules are used to further reduce the size of the matrix. Once this has been achieved, the free bond matrix is weighted according to several rules. Using HMQC or HSQC spectra in combination with an HMBC spectrum yields a one or two bond C–C connectivity (Fig. 5), and therefore the probability that two carbons, correlated in this way, are connected is 0.5. The weighting factor is scaled by this probability and so the free bond matrix represents both the possibilities and probabilities of bond formation between two components. The component with the greatest weighting factor is chosen as starting point of the structure generation process. When a connection is formed between two components the free bond matrix and weighting factors are dynamically updated. Rules eliminate potential duplication of substructures and also ensure that most likely connections between components are made first, with a view to generating the most plausible complete structure early on in the generation process. As in LSD, intelligent backtracking is used. At each stage of the structure generation, substructures are evaluated to remove chemically unlikely structures, check consistency with 2D NMR data, and in addition a simple ^{13}C chemical shift check between observed and calculated values is carried out. An innovative feature to determine whether the structure generator is heading in the right direction is by checking the rate at which 2D NMR constraints are being satisfied. In general as the generation extends towards the correct structure, the number of constraints satisfied should increase. As long as this rate of constraint satisfaction is above a predetermined value, the structure generation continues, if it falls below this level generation using this particular substructure is discontinued. This is a very powerful way to direct the structure generation process, and greatly reduces the time taken to achieve a plausible solution. The authors of CISOC–SES have also determined that the fastest route to complete structures is to generate complete skeletons first before determining the nature of the bonds (single, double, triple), and this is a direct consequence of the free bond approach used. A solution to ambiguous 2D NMR constraints that arise as a result of spectral overlap and problems created by molecular symmetry is presented in a recent implementation of CISOC–SES.[24,33] An extended encoding of the constraints and weighting described above is used to allow for ambiguity, and this is used in the free bond matrix to account for all possibilities. Examples of the use of CISOC–SES are given in the recent literature,[20,34] and a demonstration version is available for download from the worldwide web.[24]

6 Determination of stereochemistry

For large biomolecules the determination of three dimensional structure is performed by using a combination of molecular modelling and constraints derived from NOE data as well as coupling constant information.[35] The distance geometry method to determine relative stereochemistry can be applied to small molecules, and involves the generation of all possible stereoisomers, energy minimisation of these structures, and checking their consistency with the NOE data. Exhaustive generation of stereoisomers, avoiding the generation of the same structure more than once, has in all cases been achieved,[36] and the molecular modelling and distance constraints are now a part of CHEMICS.[37] In CHEMICS a rough solution conformation is determined using semi-quantitative NOE data and molecular modelling (conformational searching) in an iterative procedure. The absolute stereochemistry will still need to be determined by the use of degradative methods, auxiliary reagents, or ORD–CD.

7 Conclusions

As is evident from the discussion above, CASE programs will significantly reduce the time taken to determine the structure of complex natural products. Systems that can deal with real world problems are already available, and are likely to increase in number and improve in ability in the near future. Apart from elucidating structures these systems will also be able to help the spectroscopist in assigning chemical shift data to a proposed structure. One way in which CASE systems will be invaluable is in determining whether or not there is a single unique solution to a given problem using the given spectroscopic data. In the

cases where more than one solution is returned to the operator, these often come as a surprise, because they do not adhere to biosynthetic rules or the user's prejudices. If more than one plausible structure is produced by the system, then the spectroscopist must determine the correct one through the judicious use of reference chemical shifts and model compound data. Interaction between a spectroscopist and a CASE system will remain important in order to generate the correct structure rapidly. Therefore CASE will complement the skills of the spectroscopist, not replace them. The use of CASE systems is likely to increase in the near future, and this will enable the bottleneck so often caused by structure elucidation to be removed from the natural product drug discovery process.

8 Acknowledgements

I would like to thank those who responded to my emailed questions during the writing of this review, especially Jean-Marc Nuzillard, Kimito Funatsu and Prasanth Darba. Finally I would like to thank Joe Connolly and David Rycroft of the University of Glasgow for reading this review and making valuable suggestions.

9 References

1 M. Will, W. Fachinger and J. R. Richert, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 221.
2 http://www.wiley-vch.de/cc/si_1.html
3 U. Eggenberger and G. Bodenhausen, *Anal. Chem.*, 1989, **61**, 2298.
4 A. H. Lipkus, R. A. Nieman and M. E. Munk, *J. Magn. Reson. A*, 1993, **102**, 24.
5 P. Crews, J. Rodriguez and M. Jaspars, *'Organic Structure Analysis'*, Oxford University Press, 1998.
6 E. L. Eliel and S. H. Wilen, *'Stereochemistry of Organic Compounds'*, Wiley, 1994.
7 D. G. Corley and R. C. Durley, *J. Nat. Prod.*, 1994, **57**, 1484.
8 D. C. Chauret, T. Durst, J. T. Arnason, P. Sanchez-Vindas, L. San Roman, L. Poveda and P. A. Keifer, *Tetrahedron Lett.*, 1996, **37**, 7875.
9 R. Dunkel, C. L. Mayne, M. P. Foster, C. M. Ireland, D. Li, N. L. Owen, R. J. Pugmire and D. M. Grant, *Anal. Chem.*, 1992, 64.
10 M. Jaspars, V. Pasupathy and P. Crews, *J. Org. Chem.*, 1994, **59**, 3253.
11 G. E. Martin and R. C. Crouch, in *'Modern Methods of Plant Analysis'* vol. 15, ed. H. F. Linskens and J. F. Jackson, Berlin, 1994.
12 R. Freeman, *'Spin Choreography'*, Spektrum Academic, Oxford, 1997.
13 K. Funatsu and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 190.
14 R. E. Carhart, D. H. Smith, N. A. B. Gray, J. G. Nourse and C. Djerassi, *J. Org. Chem.*, 1981, **46**, 1708.
15 B. D. Christie and M. E. Munk, *J. Am. Chem. Soc.*, 1991, **113**, 3750.
16 J.-M. Nuzillard, *Tetrahedron*, 1991, **47**, 3655.
17 K.-P. Neidig, M. Geyer, A. Gorler, C. Antz, R. Saffrich, W. Beneicke and H. R. Kalbitzer, *J. Biomolecular NMR*, 1995, **6**, 255.
18 M. Kjaer, K. V. Andersen and F. M. Poulsen, in *'Methods in Enzymology'*, vol. 239, ed. T. L. James and N. J. Oppenheimer, San Diego, 1994.
19 M. Woodley and R. Freeman, *J. Am. Chem. Soc.*, 1995, **117**, 6150.
20 C. Peng, G. Bodenhausen, S. Qui, H. H. S. Fong, N. R. Farnsworth, S. Yuan and C. Zheng, *Magn. Reson. Chem.*, 1998, **36**, 267.
21 C. Peng, S. Yuan, C. Zheng and Y. Hui, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 805.
22 K. Funatsu, Y. Susuta and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 6.
23 K. Funatsu, N. Miyabayashi and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 18.
24 http://www.specres.com/nmrsams.html
25 http://www.mathe2.uni-bayreuth.de/molgen4/
26 I. P. Bangov, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 318.
27 I. P. Bangov, S. Simova, D. Cabrol-Bass and I. Laude, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 546.
28 K. Balasubramanian, *Chem. Rev.*, 1985, **85**, 599.
29 J.-M. Nuzillard, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 723.
30 J.-M. Nuzillard, W. Naanaa and S. Pimont, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 1068.
31 S. V. Ley, K. Doherty, G. Massiot and J.-M. Nuzillard, *Tetrahedron*, 1994, **50**, 12267.
32 G. Almanza, L. Balderrama, C. Labbe, C. Lavaud, G. Massiot, J.-M. Nuzillard, J. D. Connolly, L. J. Farrugia and D. S. Rycroft, *Tetrahedron*, 1997, **53**, 14719.
33 C. Peng, S. Yuan, C. Zheng, Z. Shi and H. Wu, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 539.
34 C. Peng, S. Yuan, C. Zheng, H. Yongzheng, H. Wu, K. Ma and X. Han, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 814.
35 J. N. S. Evans, *'Biomolecular NMR spectroscopy'*, Oxford University Press, 1995.
36 M. Razinger, K. Balasubramanian, M. Perdih and M. E. Munk, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 812.
37 K. Funatsu, M. Nishizaki and S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 645.

Review 8/04433C