Computer-Assisted Structure Elucidation: Application of CISOC–SES to the Resonance Assignment and Structure Generation of Betulinic Acid

Chen Peng,[1]*† Geoffrey Bodenhausen,[1]‡ Shengxiang Qiu,[2] Harry H. S. Fong,[2] Norman R. Farnsworth,[2] Shengang Yuan[3] and Chongzhi Zheng[3]

[1] Center for Interdisciplinary Magnetic Resonance, National High Magnetic Field Laboratory, 1800 East Paul Dirac Drive, Tallahassee, Florida 32310, USA
[2] Program for Collaboration Research in the Pharmaceutical Sciences, College of Pharmacy, 833 South Wood Street, Chicago, Illinois 60612, USA
[3] Laboratory of Computer Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Lu, Shanghai 200032, China

Received 22 May 1997; revised 24 October 1997; accepted 24 October 1997

ABSTRACT: The application of a computer-assisted structure elucidation expert system, CISOC–SES, leading to the unequivocal ^{1}H, ^{13}C and NOE resonance assignment of betulinic acid, a biologically active triterpenoid with complicated NMR resonances, is described. The procedure consists of peak picking that is independent of background information, systematic interpretation of connectivity information from 2D NMR into bond constraints and resonance assignment based on the proposed structure. *De novo* structure generation based solely on the molecular formula and spectral data is also described. This application demonstrates the potential of efficient and systematic structure elucidation of natural products with modern high-resolution NMR spectroscopy combined with artificial intelligence. © 1998 John Wiley & Sons, Ltd.

KEYWORDS: 1D NMR; 2D NMR; computer-assisted structure elucidation; betulinic acid; automated resonance assignment; automated structure generation; natural products

INTRODUCTION

Structure elucidation is one of the areas in organic chemistry that has a very long history of attempts at employing computerization technology.[1–3] In the mid-1960s, a project called DENDRAL was started at Stanford University, which was aimed at using computerized intelligence to determine the possible structures of unknown compounds from their mass spectrometric data.[4] Similar pioneering studies led to the development of programs such as CHEMICS,[5] SESAMI[6] and DARC–EPIOS,[7] which mainly used ^{1}H and ^{13}C NMR chemical shifts for constitutional structure determination. Other approaches were based on new NMR spectroscopic methods designed to provide suitable spectral information for automated interpretation.[8,9] Because of the intrinsic complexity of this task, however, most of these attempts have been of limited scope. In contrast to other steps involved in structural studies by NMR, such as data acquisition and three-dimensional structure calculation, for which computers are being extensively used, the initial step of NMR data analysis, i.e. determination of the constitution and the resonance assignment, is still done manually. To date, there have been no truly useful and widely accepted computer programs for this purpose.

CISOC–SES (Computerized Information System for Organic Chemistry–Structure Elucidation Subsystem) is an expert system designed to assist chemists in determining the constitution of unknown organic and natural product compounds with real-world complexity.[10–12] The system interprets the through-bond connectivity information (such as that from DQF-COSY and HMBC spectra) as bond constraints, i.e. requirements of a range of bond separation between the correlated spins. Such ^{1}H–^{1}H or ^{13}C–^{1}H bond constraints are mapped to ^{13}C–^{13}C ones based on the one-bond ^{13}C–^{1}H connectivities derived from the HMQC spectrum. Then it determines one or more possible sets of structural building blocks, such as —CH_3 and —CH_2—, from the molecular formula and ^{13}C/DEPT data. The structure generator of CISOC–SES searches all the possible ways to assemble the building blocks into complete structures that are compatible with the bond constraints. CISOC–SES has a unique capability to use real-world spectroscopic data that may be ambiguous and incomplete. Moreover, while the low efficiency of structure generation hampers any attempts at automated structure elucidation, the program features novel heuristic methods that significantly enhance the speed of structure generation.

* *Correspondence to*: C. Peng, Spectrum Research, LLC, 565 Science Drive, Suite A, Madison, Wisconsin 53711, USA.
† *Present address*: Spectrum Research, LLC, 565 Science Drive, Suite A, Madison, Wisconsin 53711, USA.
‡ *Present address*: Département de Chimie, Ecole Normale Supérieure, 24 Rue Lhomond, F-75231 Paris cedex 05, France.
Contract/grant sponsor: National High Magnetic Field Laboratory.

In previous tests of this program for its potential to solve structures of complex natural products, we employed NMR data reported in the literature, which were to a greater or lesser extent devoid of ambiguities and artifacts.[12] In addition to *de novo* structure generation, we have also developed a new function that helps chemists to evaluate a structural proposal by unequivocal 1D and 2D NMR resonance assignment. The reason is that sometimes one may propose a limited number of possible structures based on the background information. In such a case, resonance assignment based on the proposed structure is more efficient than structure generation. To evaluate the utility of this new function, the use of experimental NMR data is desirable. Betulinic acid, a biologically active compound isolated in our laboratories was chosen as the model compound not only because of its promising anti-HIV[13] and anti-melanoma[14] potential, but also because of its apparently challenging NMR complexity due to low functionality with many methylene proton signals falling into a narrow chemical shift region. This compound has been studied previously at 300 MHz and the ^{13}C resonances were assigned with much uncertainty based on comparison with reference compounds.[15,16] Only partial ^{1}H resonance assignments were reported because of severe peak overlap.[15] This paper describes the application of CISOC–SES to the structure elucidation of betulinic acid by two approaches: complete and unequivocal ^{1}H and ^{13}C resonance assignments based on a structure proposed by the user and *de novo* structure generation based solely on spectroscopic evidence. Stereospecific assignments of some pro-chiral ^{13}C and ^{1}H resonances were accomplished manually based on the constitutional structure and NOE assignment.

RESULTS AND DISCUSSION

Betulinic acid (Fig. 1) was isolated from the bark of American white-barked birches (*Betula papyrifera* Marsh., Betulaceae) and exhibits a molecular formula $C_{30}H_{48}O_3$ ($M_r = 456$) based on ^{13}C and DEPT NMR and electron ionization mass spectrometric data. The identity of our sample with betulinic acid was confirmed by comparing its physical and NMR spectral data with those reported in the literature.[15,16]

With CISOC–SES, the structure elucidation of betulinic acid consists of the following steps: (1) peak picking without reference to any proposed structure, (2) systematic interpretation of 2D NMR-derived connectivity information into bond constraints and (3) user-proposed structure-oriented resonance assignment or *de novo* structure generation based solely on the spectral information and molecular formula.

Peak Picking

The ^{1}H-decoupled ^{13}C spectrum showed 29 peaks. In Fig. 2 the peaks are numbered from high to low frequency and the chemical shifts are listed in Table 1. The multiplicities were determined from DEPT experiments. Based on its height, plus the fact that it was correlated to two methyl ^{1}H chemical shifts in HMQC [peaks labeled Q30 and Q31 in Fig. 3(a)], ^{13}C peak 27 was assigned to two degenerate methyl carbons. Hence it was taken as two peaks with slightly different ^{13}C chemical shifts (peaks 27 and 30 in Table 1).

Even at 720 MHz, severe peak overlap makes it impossible to resolve all the ^{1}H peaks in the 1D ^{1}H spectrum. Therefore, the ^{1}H chemical shifts were determined from the coordinates of the ^{13}C-decoupled HMQC cross peaks, which were well resolved. In analogy to the ^{13}C peaks, these 1D ^{1}H peaks were also numbered from high to low frequency, as shown in Fig. 2. The ^{1}H chemical shifts of two cross peaks [Q26 and Q24 in Fig. 3(a)] were very close (1.2107 and 1.2104 ppm). They were therefore regarded as degenerate (complete overlap) and were numbered as one peak (No. 26). Peak overlap prevented the use of integrals. Only a few easily recognized ^{1}H multiplicities were used. The remaining peaks were assigned ‘unknown’ multiplet structures. The ^{1}H peaks are listed in Table 2. Three very broad ^{1}H peaks at 4.88, 5.58 and 14.83 ppm, which apparently correspond to the water, hydroxyl and carboxyl protons, respectively, are ignored.

The HMQC peaks are also listed in Table 1. The artificially separated ^{13}C peaks 26 and 30 were arbitrarily assigned to ^{1}H peaks 28 and 31, respectively. For CISOC–SES, the peaks are input as correlated pairs of ^{13}C and ^{1}H peaks. As the HMQC spectrum is usually very clean and the number of corresponding peaks for each carbon can be predicted from its multiplicity, unambiguous peak picking is usually possible.

The cross peaks in the DQF-COSY, HMBC and NOESY spectra were identified by searching the 2D cross peaks at the 1D peak coordinates. A well resolved cross peak can be identified as connecting two correlated 1D peaks. In many cases, it was impossible to tell exactly which, among two or more overlapping 1D peaks, was the actually correlated one. CISOC–SES accepts all of the alternate 1D peaks as ‘ambiguous nodes’ and treats such a cross peak as an ambiguous connectivity. One example is the cross peak labeled C23 in Fig. 3(b). It is clear that this corresponds to H-32 in ω_1, but it is impossible to tell to which resonance, among H-21, H-22 and H-23, it corresponds in ω_2. The three ^{1}H peaks, 21, 22 and 23, which are barely separated in HMQC, are only about 0.001 ppm (less than 1 Hz at 720 MHz) away from each other, while the cross peak has a width of *ca.* 55 Hz. This ambiguous connectivity was input to CISOC–SES in the following simple format:

$$23 \cdot (21\ 22\ 23–32)$$

The first number, 23, is the ID of the cross peak. The following three numbers, 21, 22 and 23, are the IDs of the possibly correlated ^{1}H peaks in ω_2, and the last number, 32, is the ID of the correlated ^{1}H peak in ω_1. In the subsequent resonance assignment or structure

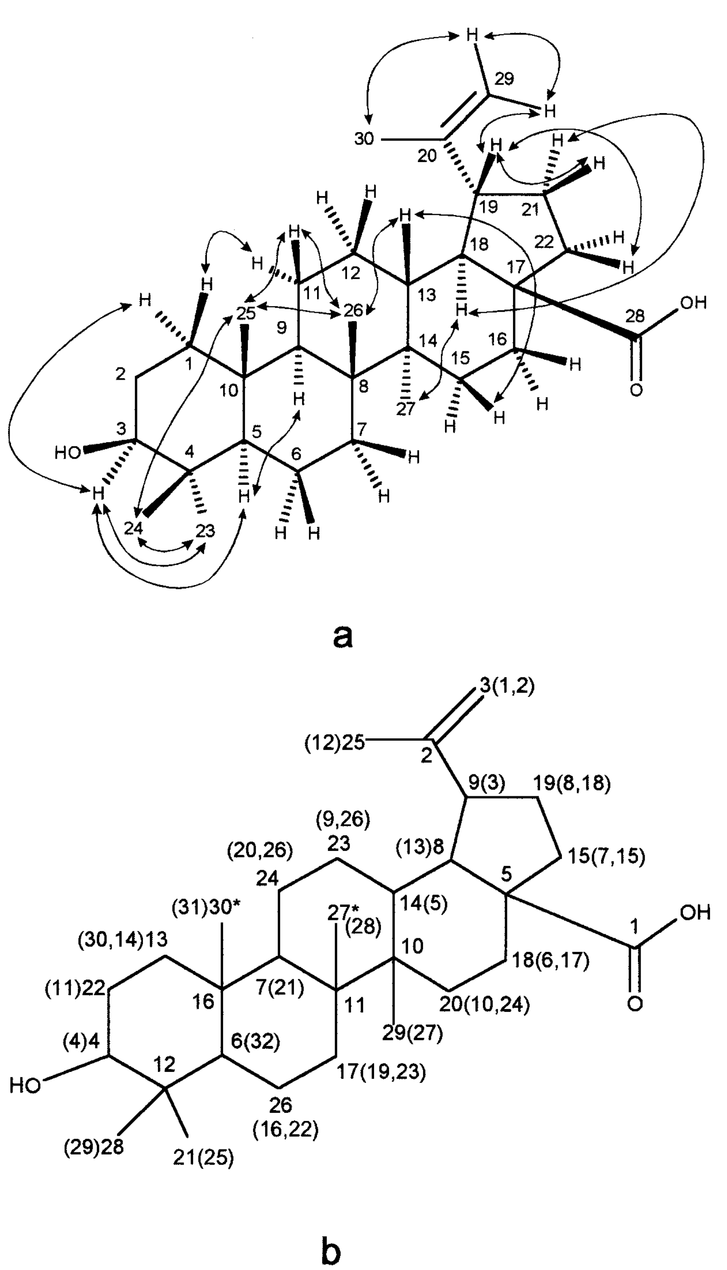

Figure 1. Structural diagrams of betulinic acid. (a) Structure proposed based on background information and with configuration deduced from NOE measurements. Key NOE connectivities are shown by arrows. The IUPAC convention is used for the numbering of the carbon atoms. (b) Topological structure given by CISOC–SES with ^{1}H and ^{13}C assignments. The numbering, in parentheses for protons and others for carbons, corresponds to that of the ^{1}H and ^{13}C peaks listed in Tables 1 and 2 and in Figs 2 and 3. The ^{13}C chemical shifts of the two methyl groups marked with asterisks are degenerate.

generation, CISOC–SES will make sure that at least one of the three possible proton pairs have a bond separation of two or three bonds in the structure. Actually, none of the cross peaks relating to the three overlapping ^{1}H chemical shifts (21, 22 and 23) could be resolved in the ^{1}H dimension, so they were all input as ambiguous connectivities.

As usual, ambiguity was even more severe in the HMBC spectrum, which contained many cross peaks and suffered from t_1 noise and low digital resolution in ω_1. Some cross peaks had ambiguous nodes in both ω_1 and ω_2 dimensions, although the correlated ^{13}C peaks were well resolved in the 1D spectrum. For example, the cross peak labeled B103 in Fig. 3(c) has three ambiguous nodes in ω_1 (C-27, C-28 and C-30) and three in ω_2 (H-21, H-22 and H-23). It is worth noting that 21 of the HMBC peaks had three ambiguous nodes, 13 had four, three had five and one had six. Occasionally, it was hard to tell a cross peak from an artifact. In that case, a probability of 0.5 was assigned to the peak. In the subsequent analysis, CISOC–SES will treat it as an unreliable correlation and allow it be violated in the resulting assignment or generated structure. For example, the HMBC cross peak labeled B78 in Fig. 3(c) appeared as a weak peak close to the t_1 ridge. It was difficult to tell whether it was a real peak or an artifact so it was input

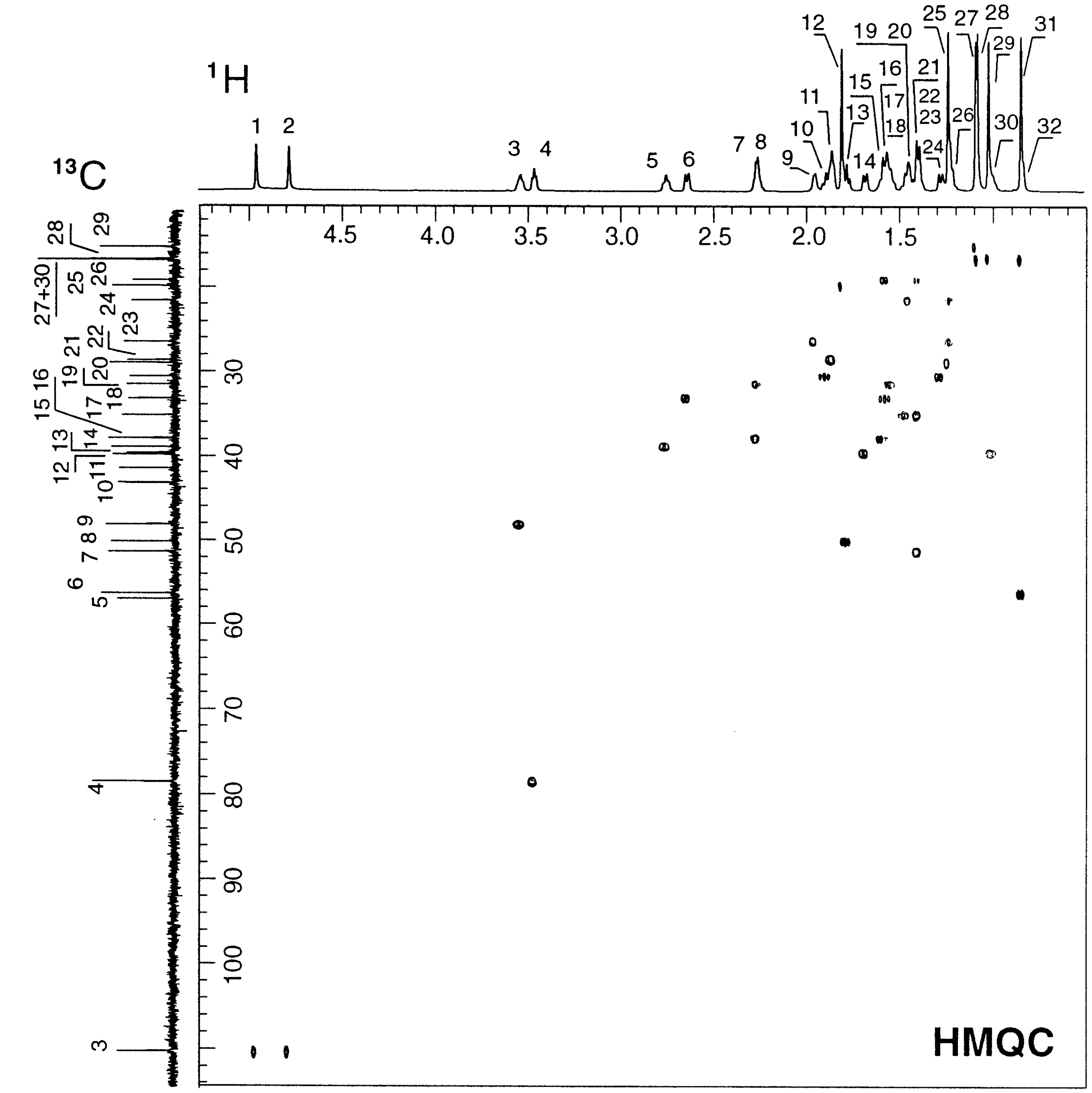

Figure 2. ^{13}C-decoupled HMQC spectrum of betulinic acid with the 1D ^{1}H and ^{1}H-decoupled ^{13}C spectra. The numbering of the ^{13}C and ^{1}H peaks is consistent with Tables 1 and 2 and Fig. 1(b). The two most downfield ^{13}C peaks, which do not have any cross peaks in the HMQC spectrum, are not displayed.

as an unreliable peak. The unsuppressed one-bond C–H peaks, which typically appear as split pairs, were discarded. Unless they coincide with the chemical shifts of other ^{1}H spins, such one-bond C–H peaks are usually readily recognized.

TOCSY peaks are not used by CISOC–SES, but the TOCSY spectrum can be used to assist the identification of DQF-COSY peaks. For betulinic acid, it was found that the TOCSY spectrum was very useful in verifying the existence of near-diagonal DQF-COSY cross peaks, which were difficult to discern in the DQF-COSY spectrum itself.

CISOC–SES allows the user to specify an 'intensity level,' which is a semi-quantitative description (i.e. strong, medium or weak) of the peak intensity, for each cross peak. Very weak DQF-COSY peaks, which probably arise from small long-range J-coupling constants, were assigned a weak intensity level. The NOESY peaks were roughly divided into three levels of intensities, i.e. weak, medium and strong, depending on their numbers of contours in the spectral plots (Table 2).

Tables 1 and 2 list the 1D chemical shifts and 2D connectivities observed for betulinic acid. For the ambiguous connectivities, which are marked by asterisks, only the real correlated chemical shifts are listed for clarity. However, it must be emphasized that, as described in subsequent sections, such ambiguous connectivities were resolved *after* the correct structure and

Table 1. ^{13}C peaks and HMQC, HMBC correlations observed for betulinic acid

^{13}C ID[a]	Chemical shift (ppm)	Multiplicity[b]	HMQC		HMBC[c]							
1 (28)	178.8	s			2.25*	1.77	1.57	1.55*	2.63			
2 (20)	151.3	s			4.95	4.77	3.52	1.79	1.77	1.53*		
3 (29)	109.9	t	4.95	4.77	3.52	1.79						
4 (3)	78.1	d	3.45		1.85*	1.67	1.22	1.00				
5 (17)	56.6	s			2.74	2.63	2.25	1.77	1.57	1.55*	1.26	
6 (5)	56.0	d	0.82		1.67	1.45	1.38*	1.22	1.00	0.83*		
7 (9)	51.0	d	1.38		1.94	1.43	1.38*	1.21	1.06	0.99	0.83	
8 (18)	49.8	d	1.77		3.52	2.74	2.63	2.25	2.24			
9 (19)	47.8	d	3.52		4.95	4.77	2.25*	1.79	1.77	1.53*		
10 (14)	42.9	s			2.74	2.63	1.94	1.77	1.45	1.26	1.07	1.06
11 (8)	41.1	s			1.56*	1.45	1.43	1.38*	1.07	1.06		
12 (4)	39.5	s			3.45	1.85	1.22	1.00*				
13 (1)	39.3	t	1.67	0.99	1.85	0.83*						
14 (13)	38.6	d	2.74		3.52	1.94	1.77	1.43	1.21	1.07		
15 (22)	37.6	t	2.25	1.57	2.24*	1.77*	1.55*					
16 (10)	37.5	s			1.85*	1.67*	1.56*	1.38*	0.83*			
17 (7)	34.9	t	1.45	1.38	1.56*	1.38*	1.06*	0.82				
18 (16)	32.9	t	2.63	1.55	1.77	1.57*	1.26					
19 (21)	31.2	t	2.24	1.53	1.57*							
20 (15)	30.3	t	1.88	1.26	1.55*	1.07						
21 (23)	28.7	q	1.22		3.45	1.00*						
22 (2)	28.3	t	1.85		1.67	0.99*						
23 (12)	26.1	t	1.94	1.21	2.74	1.77	1.21					
24 (11)	21.2	t	1.43	1.21	2.74	1.38*	1.21*					
25 (30)	19.5	q	1.79		4.95	4.77	3.52					
26 (6)	18.8	t	1.56	1.38	1.45	1.38*						
27 (26)	16.4	q	1.07		1.45*	1.38*						
28 (24)	16.3	q	1.00		3.45*	1.22*	0.82*					
29 (27)	14.9	q	1.07		2.74	1.88	1.26					
30 (25)	16.4	q	0.83		1.67*	1.38*						

[a] The IDs of the ^{13}C peaks correspond to the numbering in Fig. 1(b). The IUPAC numbering of the carbon atoms is listed in parentheses. Peaks 27 and 30 are degenerate peaks that were artificially separated.

[b] The ^{13}C multiplicities were derived from DEPT spectra: s, singlet; d, doublet; t, triplet; q, quartet.

[c] Asterisks denote ambiguous connectivities that involved other possibly correlated ^{1}H and/or ^{13}C peaks in the input data. The ambiguities were resolved based on the molecular structure and resonance assignment.

assignment had been determined starting from the ambiguous data. Of course, spectral ambiguities can be alleviated to some extent by selecting experiments such as pulse-field-gradient HMBC[17] or HSQC combined with f_1 linear prediction.[18,19] This may be important for complex molecules studied at lower magnetic fields.

Spectral Interpretation

The 1D and 2D NMR data and the molecular formula were entered into CISOC–SES via a text file. From the elemental composition and the number of ^{13}C peaks, the program recognized that the target structure must be asymmetric. From the ^{13}C multiplicities, the program assigned a certain number of protons to each carbon atom, which was labeled with the ^{13}C chemical shifts. The two residual protons were assigned to two of the three oxygen atoms. These structural fragments, each consisting a heavy atom plus a certain number of attached protons and unsatisfied valences, are used in subsequent steps as the building blocks for structure generation.

The 2D cross peaks were interpreted in terms of *bond constraints* (BCs), i.e. requirements of one or several bonds intervening between the relevant nodes (spins or atoms) in the final structure. (These bond constraints were also referred to as *topological distance constraints* in previous publications. Examples are described below.) A strong or medium DQF-COSY peak was interpreted as a BC of two or three bonds (geminal or vicinal) between the relevant protons. A weak DQF-COSY peak was interpreted as a BC of three to five bonds to include the possibility of long-range coupling. A special feature of the interpretation of DQF-COSY data is the use of 'negative' information, i.e. disconnectivity can be inferred from the absence of a cross peak. Normally, two proton-bearing carbons cannot be neighbors if no COSY peak is observed between the protons. However, this is not always true since, depending on the conformation, the scalar coupling between

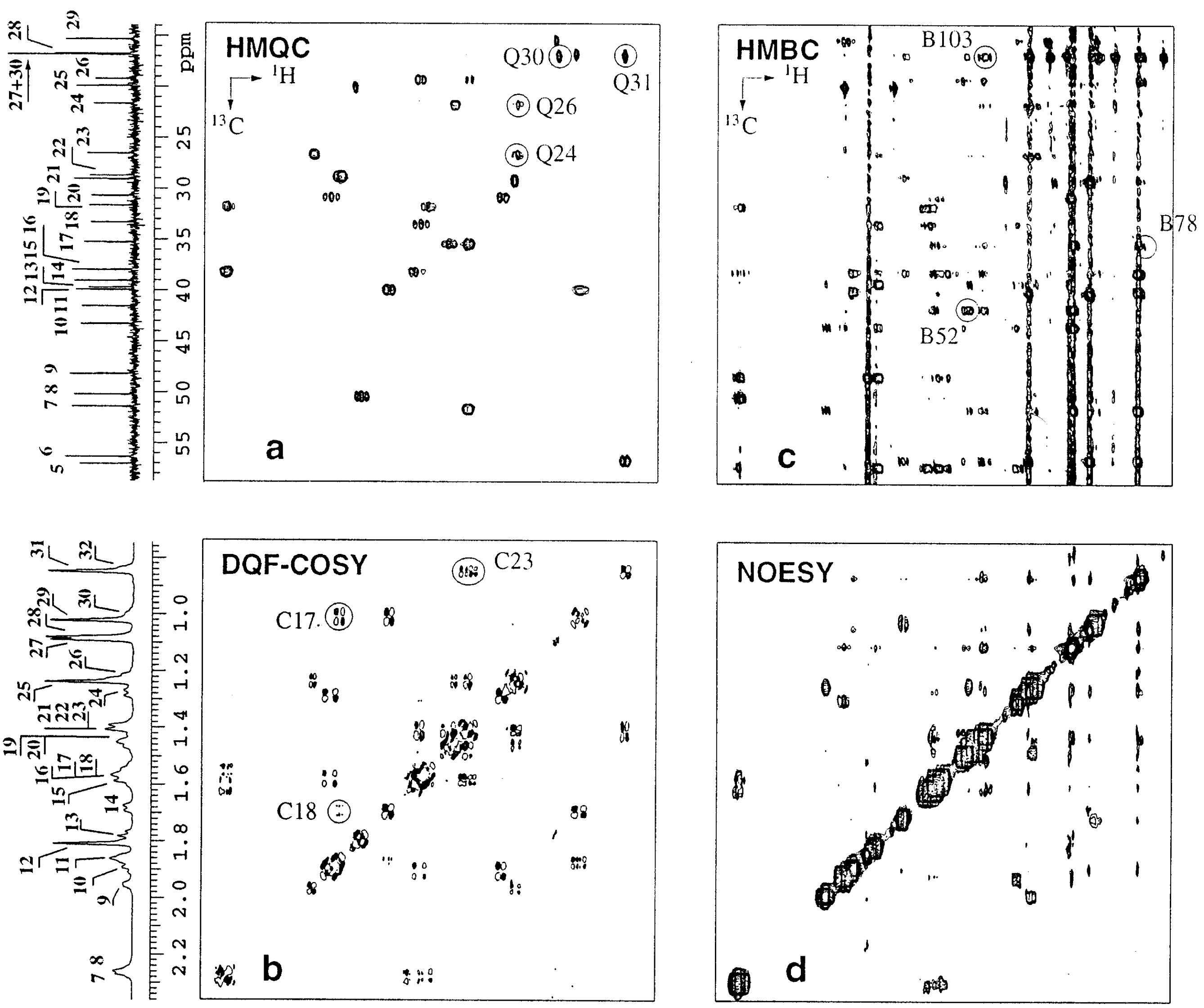

Figure 3. Upfield regions of the HMQC, DQF-COSY, HMBC and NOESY spectra of betulinic acid with the 1D ^{1}H and ^{13}C spectra. Identical ^{1}H and/or ^{13}C chemical shift scales are used for all spectra. The numbering of the ^{13}C and ^{1}H peaks corresponds to Tables 1 and 2 and Fig. 1(b). Some cross peaks are denoted by their IDs (with C, Q and B standing for COSY, HMQC and HMBC spectra, respectively).

two vicinal protons might vanish. Moreover, a non-vanishing cross peak might be neglected for practical reasons such as proximity to the diagonal, weak intensity or peak overlap. CISOC–SES provides several options to avoid these pitfalls which the user can choose at his or her discretion. We chose to use the negative information of DQF-COSY in conjunction with NOESY, i.e. two proton-bearing carbon atoms were considered to be disconnected only if neither a DQF-COSY nor a NOESY peak could be observed between the protons. Special caution was given to near-diagonal cross peaks. For this purpose, the user is prompted to add a 'pseudo bond constraint' between two protons with a chemical shift difference smaller than 0.02 ppm (14.4 Hz) in case the cross peak was overlooked. Among the 12 close pairs of ^{1}H peaks, we allowed the program to include such pseudo BCs for the ^{1}H peak pairs H-7–H-8, H-16–H-17, H-19–H-20, H-21–H-22, H-21–H-23 and H-22–H-23. For the remaining close pairs, it could be seen unambiguously in the TOCSY spectrum that no cross peaks existed so no pseudo BCs were included between them. This prevented the program from excluding possible bonds between the carbon atoms associated with these protons. The interpretation of the DQF-COSY peaks resulted in 30 ^{1}H–^{1}H BCs, including six pseudo BCs between the close pairs of ^{1}H peaks.

The HMQC cross peaks were interpreted as 33 one-bond ^{13}C–^{1}H BCs. The HMBC peaks were interpreted as 115 ^{13}C–^{1}H BCs of 2–3 bonds. For these spectra, negative information was not considered.

Based on the HMQC-derived direct ^{13}C–^{1}H connectivities, the other BCs were transformed into a homogeneous set of ^{13}C–^{13}C BCs. The resulting BCs were mutually cross-checked, and where contradictions were found, the user was warned or prompted to resolve the controversy. One example was the presence of two DQF-COSY peaks [peaks labeled C17 and C18 in Fig.

Table 2. ^{1}H peaks and DQF-COSY, NOESY correlations observed for betulinic acid[a]

^{1}H ID[b]	Chemical shift (ppm)	Multiplicity[c]	COSY[d]	NOESY[d]
1 (29α)	4.95	s	4.77 (w) 1.79 (w)	4.77 (s) 3.52 (w) 1.79* (w)
2 (29β)	4.77	s	1.79 (w)	1.79 (w)
3 (19)	3.52	m	1.77 2.24* 1.53* (w)	2.24* (w)
4 (3)	3.45	m	1.85	1.85 (w) 1.22 (m) 0.99 (w) 0.82 (w)
5 (13)	2.74	m	1.77 1.21	1.94 (w) 1.06 (m)
6 (16β)	2.63	m	1.55* 1.88 (w)	1.55* (s) 1.26 (w)
7 (22β)	2.25	m	1.57	1.57* (s) 1.55* (s)
8 (21β)	2.24	m	1.53*	1.53* (s)
9 (12β)	1.94	m	1.21	1.43 (w) 1.21 (s)
10 (15β)	1.88	m	1.55 1.26	1.26 (s) 1.06 (s)
11 (2)	1.85	m	0.99 1.67 (w)	1.67 (w) 1.00 (m) 0.83 (m)
12 (30)	1.79	s		1.53 (w) 1.21 (w)
13 (18)	1.77	m		1.55 (m) 1.07 (s)
14 (1β)	1.67	m	0.99	1.43 (w) 0.99 (s) 0.83 (w)
15 (22α)	1.57	m		
16 (6α)	1.56	m	1.38*	1.45* (m) 1.38* (s) 1.22 (m) 0.82* (w)
17 (16α)	1.55	m	1.26* (w)	1.07 (w)
18 (21α)	1.53	m		
19 (7α)	1.45	m		1.38* (s) 1.07 (w)
20 (11α)	1.43	m	1.21	1.21 (s) 1.07 (s) 0.83 (m)
21 (9)	1.38	m	1.21*	0.83* (s)
22 (6β)	1.38	m	0.82*	0.83 (s) 1.00* (m)
23 (7β)	1.38	m		1.06* (s)
24 (15α)	1.26	m		1.07 (m)
25 (23)	1.22	s		1.00 (s) 0.83 (s)
26 (11β, 12α)	1.21	m		1.07* (s) 0.83 (m)
27 (27)	1.07	s		
28 (26)	1.06	s		0.83 (s)
29 (24)	1.00	s		0.83 (s) 0.82 (m)
30 (1α)	0.99	m		0.82 (w)
31 (25)	0.83	s		
32 (5)	0.82	m		

[a] Because of severe overlap in the ^{1}H spectrum, the ^{1}H chemical shifts were measured in the HMQC spectrum. The broad ^{1}H peaks at 5.58 and 14.83 ppm in the ^{1}H spectrum, arising from the hydroxyl and carboxylic acid protons respectively, are not listed.

[b] The peak IDs correspond to the numbering of protons in Fig. 1(b). The IUPAC system numbering is listed in parentheses.

[c] s, Singlet; m, unresolved multiplets.

[d] Asterisks denote ambiguous connectivities that involved other possibly correlated ^{1}H peaks in any dimension in the input data. The ambiguities were resolved based on the molecular structure and resonance assignments. For COSY, w denotes long-range couplings. For NOESY, s, m and w denote strong, medium and weak peaks, respectively.

3(b)] concerning the protons on the same carbon pair, C-13 and C-22. The weak DQF-COSY peak C18 observed between ^{1}H peaks 11 and 14 was interpreted as a BC of 3–5 bonds as follows:

$$(11\text{–}14{:}\ 3 \sim 5;\ 0;\ 1 \sim 1)\mathrm{C18} \tag{1}$$

where the numbers before the colon represent the IDs of the correlated ^{1}H peaks in the two dimensions, the following two numbers the minimum and maximum numbers of intervening bonds between the relevant protons, next the bond type (0, 1, 2 or 3 for unknown, single, double or triple bond), and then the minimum and maximum numbers of atom pair(s) that must satisfy this constraint in the molecular structure. Appended after the closing parenthesis is the code of the cross peak from which the BC was derived. Here C18 represent COSY peak No. 18. (For a more detailed explanation of the format of BCs, see Ref. 11.)

By substituting the ^{1}H peaks with their directly connected ^{13}C peaks observed in the HMQC spectrum and appending the two relevant HMQC peaks to the peak codes, the BC in Eqn (1) was transformed into a ^{13}C–^{13}C BC as follows:

$$(22\text{–}13{:}\ 1 \sim 3;\ 0;\ 1 \sim 1)\mathrm{C18Q22Q8} \tag{2}$$

The DQF-COSY peak C17 which had a normal intensity was interpreted as a ^{1}H–^{1}H BC of 2–3 bonds as follows.

$$(11\text{–}30{:}\ 2 \sim 3;\ 0;\ 1 \sim 1)\mathrm{C17} \tag{3}$$

This BC was transformed into the following BC between the same ^{13}C spin pair as in Eqn (2):

$$(22\text{–}13{:}\ 1 \sim 1;\ 0;\ 1 \sim 1)\mathrm{C17Q22Q9} \tag{4}$$

As the bond ranges were different in Eqns (2) and (4), the program adopted the intersection of the bond separations (i.e. one bond in this case) as the final distance and obtained the following ^{13}C–^{13}C BC:

$$(22\text{–}13\colon 1 \sim 1;\, 0;\, 1 \sim 1)\mathrm{C17Q22Q9C18Q22Q8} \qquad (5)$$

If no intersection existed, the user would be prompted to supply an appropriate bond separation.

In total, 110 ^{13}C–^{13}C BCs were obtained and the first 30 BCs are listed in Table 3. It is interesting that ^{13}C peaks relevant to some of the BCs are subsets of those that occur in other BCs (e.g. BC Nos 28 and 29 in Table 3). Since the molecule is asymmetric, i.e. each ^{13}C peak corresponds to one carbon atom, the ^{13}C–^{13}C BCs can therefore be used directly as bond constraints on the carbon atoms.

As the presence of a carboxylic group was evident from the ^{13}C chemical shift of 178.8 ppm and the IR absorption, it was entered as user-known information in a similar format to the NMR-derived bond constraints.

The results obtained above were condensed into a matrix summarizing the possibilities of bond formation between the building blocks and a list of unsatisfied bond constraints. The atom–atom connectivity matrix was further reduced, weighted and reordered according to a series of heuristic rules in order to improve the efficiency of the subsequent structure generation.[10,11]

Resonance Assignment

The two-dimensional structure of betulinic acid shown in Fig. 1(a) was defined as a target structure for reso-

Table 3. ^{13}C–^{13}C **bond constraints derived from DQF-COSY, HMQC and HMBC**[a]

1	(3–25: 2 ~ 2; 0; 1 ~ 1)C2Q1Q27C3Q2Q27B13Q27B95Q1B96Q2
2	(23–24: 1 ~ 99; 0; 1 ~ 1)QSQ24Q26
3	(9–8: 1 ~ 1; 0; 1 ~ 1)C4Q7Q6B31Q7B40Q6
4	(9–15 19: 1 ~ 1; 0; 1 ~ 2)C5Q7Q11Q17B38Q11Q17
5	(9–18 19: 2 ~ 3; 0; 1 ~ 2)C6Q7Q16Q18
6	(4–22: 1 ~ 1; 0; 1 ~ 1)C7Q3Q22
7	(14–8: 1 ~ 1; 0; 1 ~ 1)C8Q10Q6B32Q10B64Q6
8	(14–24 23: 1 ~ 1; 0; 1 ~ 2)C9Q10Q26Q24B66Q26Q24
9	(18–26 18: 0 ~ 1; 0; 1 ~ 2)C10Q15Q28Q16
10	(18–20: 1 ~ 1; 0; 1 ~ 1)C11Q15Q19C15Q19Q16B81Q20
11	(19–18 19: 0 ~ 1; 0; 1 ~ 2)C13Q17Q16Q18
12	(23–24 23: 1 ~ 1; 0; 1 ~ 2)C14Q23Q26Q24B91Q26Q24
13	(22–13: 1 ~ 1; 0; 1 ~ 1)C17Q22Q9C18Q22Q8B60Q22B87Q8
14	(26 18–7 26 17: 0 ~ 1; 0; 1 ~ 6)C20Q28Q16Q5Q29Q14
15	(24–24 23: 0 ~ 1; 0; 1 ~ 2)C21Q25Q26Q24
16	(7 26 17–24 23: 1 ~ 1; 0; 1 ~ 6)C22Q5Q29Q14Q26Q24
17	(7 26 17–6: 1 ~ 1; 0; 1 ~ 3)C23Q5Q29Q14Q4B27Q5Q29Q14
18	(26 18–20: 2 ~ 2; 0; 1 ~ 2)C24Q28Q16Q20B83Q28Q16
19	(15–19: 1 ~ 99; 0; 1 ~ 1)CSQ11Q17
20	(26–18: 1 ~ 99; 0; 1 ~ 1)CSQ28Q16
21	(17–24: 1 ~ 99; 0; 1 ~ 1)CSQ13Q25
22	(7–26: 1 ~ 99; 0; 1 ~ 1)CSQ5Q29
23	(7–17: 1 ~ 99; 0; 1 ~ 1)CSQ5Q14
24	(26–17: 1 ~ 2; 0; 1 ~ 1)CSQ29Q14B98Q13
25	(1–15 19: 1 ~ 2; 0; 1 ~ 2)B1Q11Q17
26	(1–8: 1 ~ 2; 0; 1 ~ 1)B2Q6
27	(1–15: 1 ~ 2; 0; 1 ~ 1)B3Q12
28	(1–26 18: 1 ~ 2; 0; 1 ~ 2)B4Q28Q16
29	(1–18: 1 ~ 2; 0; 1 ~ 1)B5Q15
30	(2–3: 1 ~ 2; 0; 1 ~ 1)B6Q1B7Q2

[a] Only the first 30 of the total 110 bond constraints are listed. For each bond constraint, the numbers between the opening parenthesis and the colon represent the IDs of the (possibly) correlated ^{13}C spins (see Table 1 for their chemical shifts); the following two numbers the minimum and maximum numbers of intervening bonds between the relevant ^{13}C spins, next the bond type (0, 1, 2 or 3 for unknown, single, double or triple bond, respectively) and the last two numbers before the closing parenthesis are the minimum and maximum numbers of relevant spin pairs that must satisfy this constraint in the molecular structure. The appended text string designates the cross peaks from which the bond constraint is derived, where C, Q, B and N stand for COSY, HMQC, HMBC and NOESY, respectively, and the numbers are the IDs of the cross peaks. S stands for a pseudo bond constraint added by the program for near-diagonal DQF-COSY peaks. The pseudo bond constraints, such as Nos 2 and 19 in the table, have a vague bond separation of 1–99 bonds, which means that the relevant spins can be of any bonds apart in the structure.

Table 4. ^{1}H and ^{13}C resonance assignment of betulinic acid (^{1}H, 719.95 MHz; ^{13}C, 181.05 MHz)[a]

Carbon No.[a]	δ ^{13}C (ppm)	δ ^{1}H (ppm)
1	39.31	α: 0.99 (m); β: 1.67 (broad, d, J = 12.9 Hz)
2	28.32	1.85 (m)
3	78.14	3.45 (t, J = 7.2 Hz)
4	39.53	
5	55.95	0.82 (m)
6	18.81	α: 1.56 (m); β: 1.38 (m)
7	34.86	α: 1.45 (m); β: 1.38 (m)
8	41.14	
9	50.99	1.38 (m)
10	37.55	
11	21.23	α: 1.43 (m); β: 1.21 (m)
12	26.15	α: 1.21 (m); β: 1.94 (m)
13	38.65	2.74 (m)
14	42.87	
15	30.29	α: 1.26 (m); β: 1.88 (m)
16	32.89	α: 1.55 (m); β: 2.63 (m)
17	56.64	
18	49.80	1.77 (t, J = 11.5 Hz)
19	47.78	3.52 (m)
20	151.32	
21	31.24	α: 1.53 (m); β: 2.24 (m)
22	37.57	α: 1.57 (m); β: 2.25 (m)
23	28.67	1.22 (s)
24	16.33	1.00 (s)
25	16.43	0.83 (s)
26	16.43	1.06 (s)
27	14.93	1.07 (s)
28	178.82	
29	109.92	α: 4.95 (s); β: 4.77 (s)
30	19.50	1.79 (s)

[a] The IUPAC numbering of atoms is used in this table and in Fig. 1(a).

nance assignment. With CISOC–SES, an assignment matrix was set up to summarize the possible assignments of the ^{13}C chemical shifts to the carbon atoms in the target structure. The matrix was reduced based on the ^{13}C chemical shifts and multiplicity information.

The resonance assignment was carried out by systematically searching all possible ways of mapping the structural building blocks to the constituent heavy atoms in the target structure. Each possible (partial) mapping and the resulting (sub)structure was evaluated based on the assignment matrix, the connectivity matrix and the bond constraints obtained as described above. This process was completed by CISOC–SES in a few seconds and two alternative ^{13}C assignments were obtained. The two assignments arose from the interchange between ^{13}C peaks 21 and 28 to the two topologically equivalent methyl groups, which could only be discriminated based on stereochemical information.

Once the ^{13}C peaks had been assigned to the individual carbons, the assignment of the ^{1}H resonances was straightforward based on the HMQC-derived connectivities. When the ^{1}H peaks were also assigned, the 2D cross peaks were also assigned because each of them is represented by a connectivity of 1D peaks. As described previously, however, some of the COSY, NOESY and HMBC cross peaks were picked as ambiguous connectivities with more than two possibly correlated 1D peaks. These peaks needed to be further resolved. Based on the ^{1}H assignment and the actual bond separations, the ambiguous NOESY peaks were resolved by excluding impossible ^{1}H peaks. But rigorously, such ambiguities can only be resolved by considering the three-dimensional structure obtained by molecular modeling. The ambiguous DQF-COSY and HMBC peaks were also resolved in a similar fashion. The resolved peaks are marked with asterisks in Tables 1 and 2.

At this stage, CISOC–SES does not take stereochemistry into consideration, so stereospecific resonance assignments must be made manually. With the assumption that the configuration of the C-3 OH group was β-oriented, based on the knowledge of the biosynthetic background of natural triterpenoids, the relative stereochemistry of the other chiral centers could be readily determined by investigating the spatial relationship of related proton pairs with the NOESY spectrum [Fig. 3(d)]. Consequently, the non-equivalent methylene

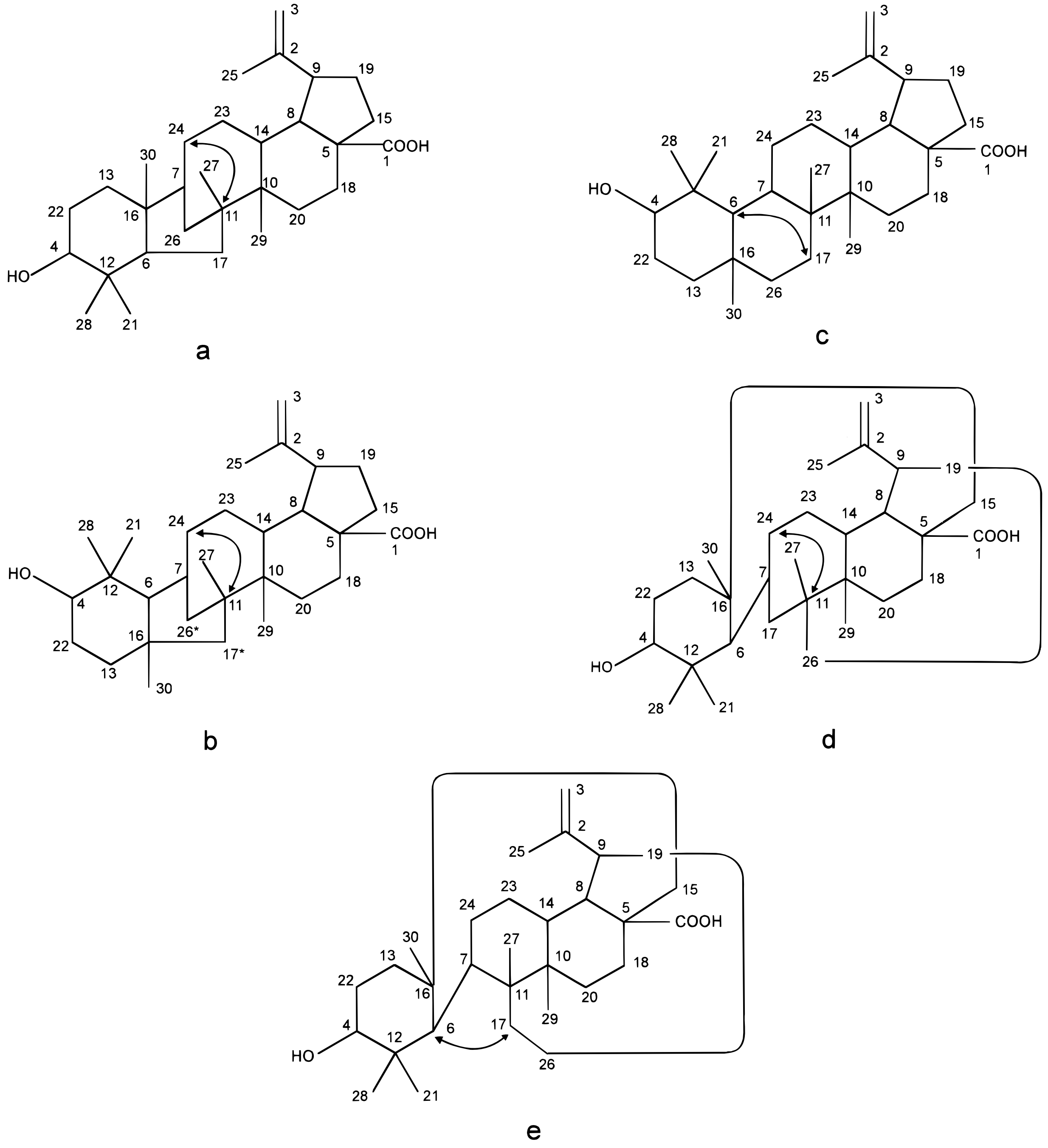

Figure 4. Five alternative structures generated by CISOC–SES for betulinic acid when the seven HMBC peaks relevant to C-7 were omitted and one bond constraint was allowed to be violated. In each structure the curved arrows indicate the pair of atoms between which the number of intervening bonds violates a known constraint of one or two bonds. Structure **b** was generated twice with interchange of the assignments of ^{13}C peaks 17 and 26 to the atoms marked with asterisks. The numbering of the carbon atoms corresponds to the IDs of the ^{13}C peaks listed in Table 1.

protons were unequivocally assigned [Fig. 1(a) and Table 4].

De novo Structure Generation

Although the proposed structure was verified by the resonance assignment of CISOC–SES, it is interesting to test the structure generation function of CISOC–SES assuming that no user-proposed structure was available. After interpreting the spectral data as bond constraints and setting up the atom–atom connectivity matrix, CISOC–SES generated the correct structure uniquely in about 1 min. The same resonance assignments were obtained as from the target structure-based resonance assignment.

Compared with target structure-based resonance assignment, one of the advantages of structure generation is that it provides important information about whether the proposed structure is the only possibility that fits the experimental data. If multiple candidate structures are obtained, it helps the user either revise the peak picking results, add new experimental data or correct the proposed structure. For example, we once missed the seven HMBC peaks of C-7 in Table 1. Using that data set, six candidate structures were generated in a CPU time of about 12 min when up to one bond constraint was allowed to be violated. The most plausible proposal, which was the only one that did not violate any bond constraints, coincided with the proposed structure [Fig. 1(b)]. The ^{13}C and ^{1}H assignments of this structure were the same as determined above. In each of the remaining five structures, shown in Fig. 4, one bond constraint was violated. These candidate structures were eliminated by confirming the existence of the cross peak(s) from which the violated BCs were derived. For example, structures **a**, **b** and **d** in Fig. 4 violated the following BC:

$$(11\text{–}24{:}\ 1 \sim 2;\ 0;\ 1 \sim 1)\mathrm{B52Q25} \qquad (6)$$

On re-examination of the HMBC spectrum, the existence of peak B52 [Fig. 3(c)], which was interpreted as a constraint of one or two bonds between carbon atoms 11 and 24, was confirmed. Of course, the unusual carbon skeletons of the alternative candidate structures could also be easily eliminated from the biosynthetic point of view.

Study of the alternate structures also led to the discovery of the book-keeping error of missing the HMBC peaks related C-7. When comparing candidate structure **a** in Fig. 4 with the correct structure [Fig. 1(b)], it was obvious that an HMBC peaks between H-28 and C-7 would have eliminated structure **a**. The absence of the seven HMBC peaks related to C-7 were thus discovered. After adding these peaks, the correct structure was generated uniquely in a shorter computational time.

CONCLUSION

In this study, the ^{1}H and ^{13}C resonances of betulinic acid were unequivocally and completely assigned for the first time. In a previous study, only 11 of the ^{1}H peaks, most of which belonged to methyl groups, were assigned because of severe peak overlap.[15] Complete ^{13}C resonance assignments were reported in two previous studies by comparison of the assignment of similar compounds.[15,16] According to our results, the ^{13}C assignments of C-18 and C-19 were erroneously interchanged in both earlier reports.

Generally, the determination of the covalent structure and the resonance assignment are prerequisites for a detailed study of the solution structure and dynamics of natural products. For complex molecules, this process could be error prone and time-consuming. While experience and intuition are a major resource for this task by human experts and usually quickly lead to the correct structure, it is not uncommon that human biases could lead to erroneous conclusions. This paper demonstrates that with CISOC–SES, the efficiency of structure determination and resonance assignment can be significantly enhanced. Moreover, the systematic way of analyzing NMR data introduced in this paper, including the peak picking that is independent of background information and the interpretation of bond constraints derived from connectivity information prior to structure generation and resonance assignment, help to guarantee less biased structure elucidation with complete consideration of all structural possibilities. It is expected that the idea of prejudice-free and reference structure-independent structure elucidation of natural products based on 2D NMR spectroscopy[20,21] will be fully realized with the assistance of CISOC–SES.

EXPERIMENTAL

A sample of 20 kg of the bark of American white birches (*Betula papyrifera* Marsh., Betulaceae) was collected by Mr S. Totura in November 1994 and identified by Dr W. Hess of the Morton Arboretum, Lisle, IL, USA. A voucher specimen was retained at the University of Illinois, Pharmacognosy Field Station, Doreners' Grove, IL, USA. The bark was extracted with $CHCl_3$ and the extract was concentrated under vacuum to yield copious amounts of betulin (*ca.* 4 kg). The mother liquid was subjected to repeated column chromatography on silica gel to give crude betulinic acid. Recrystallization from MeOH afforded an analytically pure sample (56 mg, yield 0.028%) in the form of colorless needles, m.p. 290–292 °C, $[\alpha]_D + 7.5°$ (*c* 0.5, pyridine).

A sample of 23.8 mg of betulinic acid was dissolved in 0.7 ml of pyridine-d_5 (99.94% D; Cambridge Isotope Laboratories product). The 0.084 M solution was subjected to four freeze–pump–thaw cycles and the vessel was filled with nitrogen before being sealed. All spectra were acquired with a 720 MHz Varian Unity-Plus spectrometer at the National High Magnetic Field Laboratory. A triple-resonance inverse probe was used. The temperature was set to 30 °C for all experiments. Standard pulse sequences of VNMR version 5.1 were used for all experiments. ^{1}H spectrum was obtained using a spectral width of 12 698 Hz with 16 384 points, zero filled to 32 768, and ^{13}C spectrum was obtained using a spectral width of 39 761 Hz with 32 768 points, zero filled to 65 536. The DQF-COSY spectrum was obtained using 256 increments (16 transients per increment) and 1024 data points at a spectral width of 3680 Hz for both dimensions, zero-filled to 512 × 2048 points in ω_1 and ω_2, respectively. Using similar parameters, the NOESY spectrum was obtained using a mixing time of 200 ms. The phase-sensitive HMQC spectrum used a ^{1}H spectral width of 4284 Hz, a ^{13}C spectral width of 39 676 Hz, 1024 data points with zero filling to 2048, 256 time increments with zero filling to 512, 16 transients per increment, a relaxation delay of 1.0 s and a BIRD pulse nulling delay of 0.36 s. The HMBC spectrum had similar acquisition parameters except that 32 transients were collected per increment and the BIRD pulse and nulling delay were eliminated. Delays were optimized for $J = 8$ Hz. Chirp-95 decoupling[22] was used for both ^{1}H and ^{13}C decoupling.

All processing used standard Varian software (5.1 release). The ^{1}H and ^{13}C chemical shifts were calibrated by reference to the center peaks of pyridine at 7.55 and 135.5 ppm, respectively.

CISOC–SES is run on a Silicon Graphics UNIX workstation (SGI Indy). Currently CISOC–SES is a command-driven program. All intermediate and final results are written into a text file. 1D and 2D spectral data were peak-picked by hand and input to CISOC–SES as a text file. The user-proposed structure was entered through simple commands specifying the connectivity between the atoms. The generated structures were output as atom–atom connection tables together with the assignments of ^{13}C chemical shifts.

Recently, CISOC–SES has been further improved with a graphical user interface and combined with a program called SpecMan, which has the advanced peak picking procedures required for computer-assisted spectral analysis and structure elucidation. The new version of CISOC–SES is now called NMR–SAMS, and is available from Spectrum Research, LLC, along with the program SpecMan for advanced peak picking and interactive spectral analysis.

Acknowledgements

This research was supported by the National High Magnetic Field Laboratory, Tallahassee, FL, USA. C. Peng is grateful to Dr Prashanth Darba for helpful discussions on the manuscript.

REFERENCES

1. N. A. B. Gray, *Computer-Assisted Structure Elucidation.* Wiley, New York (1986).
2. D. H. Smith (Ed.), *Computer-Assisted Structure Elucidation. ACS Symposium Series.* American Chemical Society, Washington, DC (1977).
3. C. Peng, S. Yuan, C. Zheng and L. Chen, *Computers and Applied Chemistry. Computer Chemistry Monograph Series 4*, p. 26. Science Press, Beijing (1995).
4. R. K. Lindsay, B. G. Buchanan, E. A. Geigenbaum and J. Ledergerg, *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project.* McGraw-Hill, New York (1980).
5. S.-I. Sasaki and Y. Kudo, *J. Chem. Inf. Comput. Sci.* **25**, 252 (1985).
6. B. D. Christie and M. E. Munk, *Anal. Chim. Acta* **200**, 347 (1987).
7. J.-E. Dubois and Y. Sobel, *J. Chem. Inf. Comput. Sci.* **25**, 326 (1985).
8. U. Eggenberger and G. Bodenhausen, *Anal. Chem.* **61**, 2298 (1989).
9. P. Pfåndler and G. Bodenhausen, *Magn. Reson. Chem.* **26**, 888 (1988).
10. C. Peng, S. Yuan, C. Zheng and Y. Hui, *J. Chem. Inf. Comput. Sci.* **34**, 805 (1994).
11. C. Peng, S. Yuan and C. Zheng, *J. Chem. Inf. Comput. Sci.* **35**, 539 (1995).
12. C. Peng, S. Yuan, C. Zheng, Y. Hui, H. Wu, K. Ma and X. Han, *J. Chem. Inf. Comput. Sci.* **34**, 814 (1994).
13. T. Fujioka, Y. Kashiwada, R. E. Kilkuskie, L. M. Cosentino, L. M. Ballas, J. B. Jiang, W. P. Janzen, J.-S. Chen and K.-H. Lee, *J. Nat. Prod.* **58**, 243 (1994).
14. E. Pisha, H. Y. Chai, I.-S. Lee, T. E. Chagwedera, N. R. Farnsworth, G. A. Cordell, C. W. W. Beecher, H. H. S. Fong, A. D. Kinghorn, D. M. Brown, M. C. Wani, M. E. Wall, T. J. Hieken, T. K. D. Gupta and J. M. Pezzuto, *Nature Med.* **1**, 1046 (1995).
15. S. Siddiqui, F. Hafeez, S. Begum and B. S. Siddiqui, *J. Nat. Prod.* **51**, 229 (1988).
16. M. Sholichin, K. Yamasaki, R. Kasai and O. Tanaka, *Chem. Pharm. Bull.* **28**, 1006 (1980).
17. P. L. Rinaldi and P. A. Keifer, *J. Magn. Reson., Ser. A* **108**, 259 (1994).
18. W. F. Reynolds, S. McLean, L.-L. Tay, M. Yu, R. G. Enriguez, D. M. Estwick and K. O. Pascoe, *Magn. Reson. Chem.* **35**, 455 (1997).
19. W. F. Reynolds, M. Yu, R. G. Enriquez and I. Leon, *Magn. Reson. Chem.* **35**, 505 (1997).
20. H. Duddeck and W. Dietrich, *Structure Elucidation by Modern NMR, A Workbook.* Springer, New York (1989).
21. D. S. Rycroft, in *Studies in Natural Products Chemistry*, edited by Atta-ur-Rahman, Vol. 9, pp. 93–107. Elsevier, Amsterdam (1991).
22. R. Fu and G. Bodenhausen, *Chem. Phys. Lett.* **245**, 415 (1995).