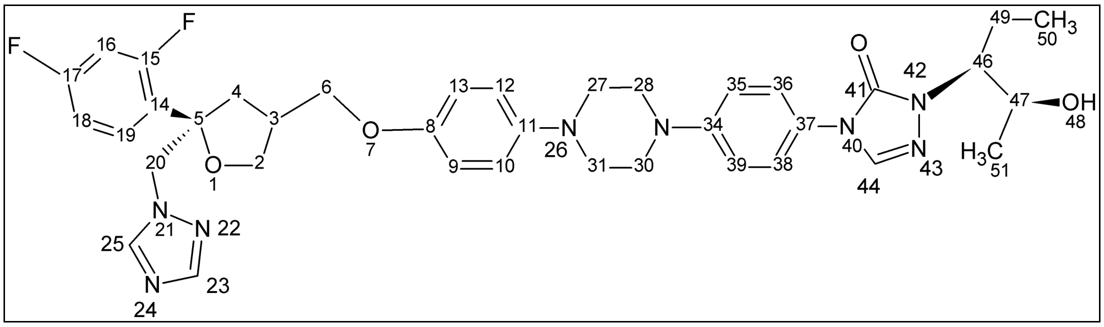
The Impact of Long-Range ^{1}H-^{15}N Heteronuclear Shift Correlation Data on Computer-Assisted Structure Elucidation: Posaconazole

Bruce D. Hilton* and Gary E. Martin

Merck Research Laboratories, Discovery and Preclinical Sciences, Process Chemistry-Rapid Structure Characterization Laboratory, Summit, NJ 07901
*E-mail: zymic56@gmail.com; lighthousephoto@gmail.com

Received August 18, 2010
DOI 10.1002/jhet.803
View this article online at wileyonlinelibrary.com.



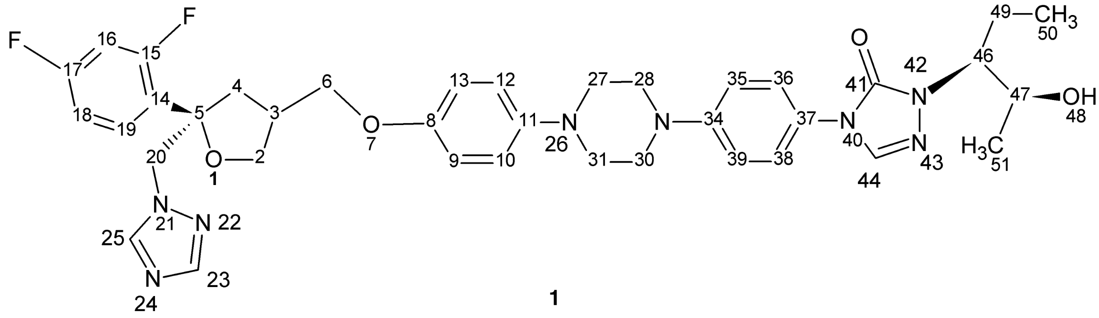
Considerable work has been invested in the area of computer-assisted structure elucidation (CASE) methods. As NMR techniques have been developed that provide more effective atom-to-atom connectivity information, it has become theoretically possible to do *de novo* structure elucidation based on 2D NMR datasets recorded for an unknown molecule. However, as annular (ring) nitrogen atoms become more prevalent in complex chemical structures, the ability to rely solely on ^{1}H and ^{13}C homo- and hetero-nuclear direct and long-range connectivity information to solve a structure correspondingly diminishes. Hence, we now wish to report the results of an investigation into the application of CASE methods with and without long-range ^{1}H-^{15}N data using posaconazole as a model compound, which has eight annular nitrogens in its structure. With the inclusion of ^{1}H-^{15}N data long-range data, the structure could be successfully determined in a few hours. Excluding the ^{1}H-^{15}N data caused the program to generate millions of candidate structures, none of which fit the data well enough to be stored.

J. Heterocyclic Chem., **49**, 526 (2012).

INTRODUCTION

The utility of obtaining long-range ^{1}H-^{15}N 2D NMR data to support the structure elucidation of polyaza drugs, alkaloids, and related molecules is intuitively obvious and has been reviewed multiple times [1–6]. Similarly, there have been several extensive recent reviews of computer-assisted structure elucidation (CASE) methods [7,8]. There have been several examples published in which authors have evaluated the performance of various CASE programs for the determination of chemical structures with and without the inclusion of long-range ^{1}H-^{15}N correlation data [9–13]. Cheatham *et al.* [14] have also recently examined the impact of 1,1-ADEQUATE and H2BC data on CASE methods but did not incorporate long-range ^{1}H-^{15}N 2D NMR data in that study. The reports that have appeared thus far, however, have not involved molecules with large number of annular (ring) nitrogen atoms incorporated in their structures, and the molecules were still successfully determined without the long-range ^{1}H-^{15}N heteronuclear correlation data albeit less efficiently [9–13]. For this reason, we were interested in exploring the potential impact of the inclusion or exclusion of long-range ^{1}H-^{15}N heteronuclear correlation data when the molecule under examination involves a relatively large number of annular nitrogens. For this reason, we elected to utilize the antifungal agent posaconazole, **1**, whose ^{15}N chemical shift assignments and long-range ^{1}H-^{15}N correlations have been recently reported [15].

The study reported here utilized the Structure Elucidator™ program v.12.01 developed by ACD Labs. Posaconazole, **1**, has a molecular formula of $C_{37}H_{42}F_{2}N_{8}O_{4}$ incorporating eight nitrogen atoms in three of the seven cyclic moieties of the structure [15]. When the data, including long-range ^{1}H-^{15}N heteronuclear correlations were used as input to the Structure Elucidator program, the program produced 16,550 structures, 87 of which passed the structure filter. The correct structure was the best fit; structure generation ran for just over 6 h. In contrast, a repeat run excluding the long-range ^{1}H-^{15}N correlation data generated over 104 million structures in



1

25 h of computing time; none of the structures fit the NMR data well enough to be stored and, at that point, the computation was stopped. The assignments of the ^1H, ^{13}C, and ^{15}N chemical shifts of posaconazole, **1**, are summarized followed by a discussion of a general approach to using CASE methods for structure elucidation and a discussion of the results obtained using the Structure Elucidator CASE program in particular.

NMR SPECTROSCOPY

A sample was prepared by dissolving 5.2 mg of posaconazole in 200 μL DMSO-d_6 (CIL) after which the sample was transferred to a 3-mm NMR tube (Wilmad) using a flexible Teflon™ needle and a Hamilton gastight syringe. All spectra were recorded with the sample temperature regulated at 25°C. Proton and carbon reference spectra, a ^1H-^{13}C HSQCAD spectrum and an 8-Hz optimized ^1H-^{13}C GHMBCAD spectrum were obtained using a Varian dual channel 500-MHz NMR spectrometer. An 8-Hz optimized ^1H-^{15}N GHMBCAD spectrum was obtained using a 300-μg sample of **1** dissolved in 30 μL of d_6-DMSO using a Bruker 600 MHz NMR spectrometer equipped with a 1.7-mm TCI gradient triple resonance MicroCryoProbe™ [15].

ASSIGNMENT OF THE NMR SPECTRA OF POSACONAZOLE

The NMR of posaconazole and several degradation products has been previously reported but resonance assignments were not detailed in that report [16]. Hence, a complete assignment was performed manually using a conventional approach. The numbered structure (**1**) is shown above; the resonance assignments are collected in Table 1. A total of 20 usable, long-range ^1H-^{15}N correlations were observed in the ^1H-^{15}N GHMBC spectrum of posaconazole [15]. The stereochemistry was not reconfirmed but has been well established. The assignment process was straightforward with only two methine resonances, those for H46 and H47, in close proximity, resonating at 3.81 and 3.80 ppm, respectively. There was, however, no ambiguity in the interpretation of the 2D spectra because of a noticeable offset of the cross peaks in the proton dimension.

STRUCTURE ELUCIDATION METHOD

In any structure elucidation based on CASE methods, contradictions can be introduced for a variety of reasons. Examples of sources of errors include: resonance marking errors and/or ambiguities in marking peaks' positions; errors caused by the user placing constraints that are too stringent on certain correlations; and errors due to mistakenly assigning resonance multiplicities in HSQCAD data among others.

The overt choice was made not to use automatic correction of inconsistencies or to allow "fuzzy generation" [7] to ensure that all the data marked were used and that all data were consistent. These choices perhaps involve some time penalty in terms of data preparation, but this approach also provides a "clean" dataset and, thereby, increased confidence in the results of the structure generation. The molecular connectivity diagram (MCD) for posaconazole is shown in Figure 1. The steps used to prepare the data for the computation run can be outlined as follows. Note that the procedure described below is sensible when one is looking to correct the raw data and/or to search for alternative structures that may fit the available data, for example, in the case of a by-product of a reaction or a degradation product. When dealing with a true unknown, one obviously cannot follow this procedure to the full structure, but the general approach still shows how the data can be thoroughly checked and plausible substructures consistent with all of the data constructed using the Structure Elucidator program package.

1. Acquire and process all 1D and 2D spectra.
2. Reference all spectra and align the 1D spectra with the 2D spectra. The alignment process, as far as it can be taken, is crucial. Successful CASE runs require that the same ^1H, ^{13}C, and ^{15}N chemical shifts are used to mark cross peaks in all of the 2D spectra.
3. Provide heteroatom information as possible. For example, there is a noncarbon bound proton resonating at 4.68 ppm based on the HSQCAD spectrum. The

Table 1

Summary of the ^{1}H, ^{13}C, and ^{15}N NMR assignments for posaconazole (**1**) in d_6-DMSO.

Assignment	δ ^{1}H (ppm)	Multiplicity, J_{HH} (Hz), J_{HF} if noted	δ ^{13}C/^{15}N, ppm	Multiplicity, J_{CF} (Hz)	ACD™ ^{13}C predictions
2	4.02, 3.74	dd, 8.7, 7.4 cm	69.9	s	70.1
3	2.53	cm	38.4	s	40.0
4	2.40, 2.13	ddd, 13.0, 8.2, 2.3 (J_{HH} or J_{HF}); dd, 13.2, 8.1	37.6	d, 2.7	38.9
5	–	–	83.3	d, 3.9	85.7
6	3.75, 3.67	dd, 9.4, 7.7 cm	68.7	s	70.5
8	–	–	152.2	s	154.2
9/13	6.80	d, 9.1	115.0	s	114.6
10/12	6.94	d, 9.2	117.6	s	116.1
11	–	–	145.4	s	144.3
14	–	–	126.2	dd, 12.8, 3.5	123.0
15	–	–	158.7	dd, 246.4, 12.4	160.8
16	7.27	2.6 cm	104.5	t, 26.5	104.2
17	–	–	161.9	dd, 246.0, 12.1	163.9
18	6.99	ddd, 8.4, 8.4 (J_{HF}), 2.4	111.0	dd, 20.6, 3.5	110.7
19	7.29	6.9 cm	128.5	dd, 9.7, 5.8	128.6
20	4.60, 4.56	d, 14.6; d, 14.6	55.2	d, 3.5	56.0
N21	–	–	212.8	s	–
N22	–	–	298.8	s	–
23	7.78	s	150.5	s	152.1
N24	–	–	252.6	s	–
25	8.34	s	145.0	s	142.8
N26	–	–	60.8	s	–
27/31	3.16	cm	49.6	s	49.9
28/30	3.31	cm	48.3	s	50.2
N29	–	–	65.4	s	–
34	–	–	149.7	s	147.2
35/39	7.10	d, 9.2	115.8	s	115.0
36/38	7.51	d, 9.1	122.7	s	125.0
37	–	–	125.7	s	123.8
N40	–	–	155.0	s	–
41	–	–	152.4	s	151.1
N42	–	–	174.2	s	–
N43	–	–	262.8	s	–
44	8.33	s	134.8	s	134.1
46	3.80	cm	62.5	s	63.9
47	3.81	cm	67.1	s	68.0
48-OH	4.68	d, 5.0	–	–	–
49	1.70	cm	21.3	s	22.2
50	0.74	t, 7.3	10.6	s	11.6
51	1.12	d, 6.0	19.9	s	19.8

^{1}H-^{15}N-data further indicates that this proton is not directly bound to nitrogen. Therefore, it must be a hydroxyl proton, which can be introduced as program input by creating a pseudo ^{1}H-^{17}O HSQC 2D spectrum.

4. Where possible, account for obvious structural features by the addition of user constraints. For example, the ^{19}F-^{13}C couplings observed in the 1D ^{13}C spectrum were used with other data to assemble the *m*-difluorophenyl ring of posaconazole.
5. Account for symmetry where possible. Although the recognition of symmetry is quite quick for an experienced spectroscopist, establishing symmetry, for example, for a 1,4-disubstituted phenyl ring is a much more time-consuming process for computer programs.
6. Conduct a consistency check of the data. In Structure Elucidator, this consistency check examines the MCD. If the consistency check fails, no structures can be generated, and the data will have to be rechecked to remove the inconsistency.
7. Next, working back to the correct set of original constraints (step 4), any user constraints added *ad hoc* are gradually removed, and the structure generator is run after each set of constraints is removed.

For the present example, the iterative, stepwise process just described and allowed us to arrive at a robust set of data and constraints that could produce the correct structure in a reasonable amount of time. Once this was achieved, it was possible to examine closely the effect of ^{15}N-related constraints as well as other groups of constraints on the structure generation, without having to do further evaluations of the individual constraints.

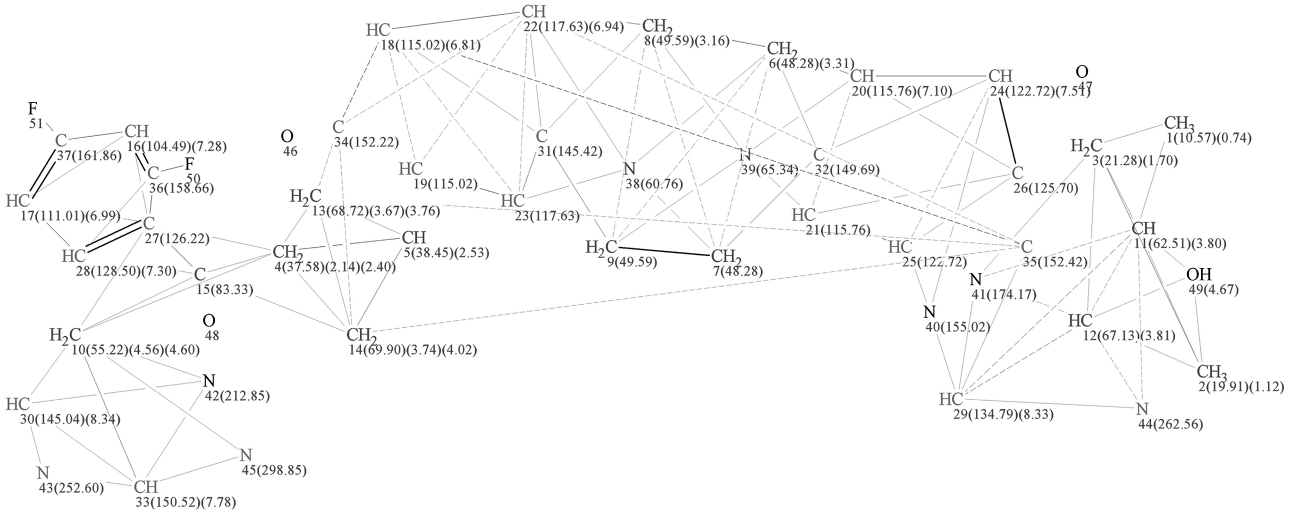


Figure 1. Molecular connectivity diagram (MCD) for posaconazole incorporating the long-range ^{1}H-^{15}N constraints (see text), Elucidator v 12.01.

STRUCTURE ELUCIDATION OF POSACONAZOLE USING STRUCTURE ELUCIDATOR

There have been numerous reports that have appeared in the literature that have treated the general topic of CASE of complex natural product structures. For a comprehensive treatment of the work that has been reported, the interested reader is referred to two recent reviews [7,8]. We will not delve into the many nuances of CASE methods, but the basic premises of this study are noted below.

The 1D proton spectrum was used to make important assignment decisions, but the spectrum itself was not included in the structure elucidation run. Our experience has shown that if all proton signals are accounted for (*e.g.,* including the OH signal discussed next) by some means, including the raw 1D ^{1}H NMR spectrum usually only serves to make it more difficult to "shape up" the data properly because of the multiplet structures arising from homonuclear coupling. A proton spectrum with integration or with full multiplet analysis creates conflicts with the 2D data, because the program does not know how to choose a consistent 1D chemical shift from the spectrum. A proton spectrum specifically peak picked to choose a single chemical shift for each proton multiplet might work, but as the location of the multiplets in the case of inevitable overlap generally comes from the 2D data, it seems more reasonable to have the program determine those chemical shifts in that manner.

In any structure elucidation using CASE methods, there is a high probability that contradictions will be created for various reasons, for example, because of mismarking of peaks, or because the program has placed constraints that are too stringent on some correlations, or because the multiplicity information in the HSQCAD spectra has been entered incorrectly in the program input. For example, in the latter case, it is usually necessary to indicate to the program that the HSQCAD spectrum is multiplicity-edited and to subsequently indicate specifically which carbons are methyls and methines. Any error in this procedure will result in a failure of the MCD during checking. In general, a very careful line-by-line review of all of the data is necessary. We chose not to use automatic correction of inconsistencies or to allow "fuzzy generation" to assure that all of the data marked are used and consistent. These choices impose a time penalty on the investigator, but they also result in a clean dataset and, thereby, increased confidence in the output of the program.

Finding a problematic correlation can be a challenge. The procedure detailed above has been found to be the best method to identify problematic correlations. In the present case, the molecule had to be almost completely constructed before the final problem correlation was uncovered; it turned out to be a GHMBCAD correlation, the constraint for which needed to be loosened from 1–3 bonds to 1–4 bonds. General loosening of all of the constraints would be impractical for a molecule of the size of posaconazole because of the consequent, drastic increase in computing time for structure elucidation that would result. While the general procedure just outlined may appear to be cumbersome, it does provide an effective means of identifying a problematic correlation or correlations.

Once the data appeared to be well conditioned all *ad hoc* user constraints were removed, and a full elucidation run was initiated using the molecular connectivity diagram shown in Figure 1. In all, 29 GCOSY, 26 ^{1}H-^{13}C HSQCAD, 20 ^{1}H-^{15}N GHMBCAD, and 78 ^{1}H-^{13}C GHMBCAD correlations were used to generate constraints for the program. One-bond user constraints were limited to six for the difluoro-phenyl ring system (see above), six for pairs of vicinally coupled protons from the two para-substituted aromatic rings and the piperazine ring (see above), and eight one-bond correlations added from symmetry considerations

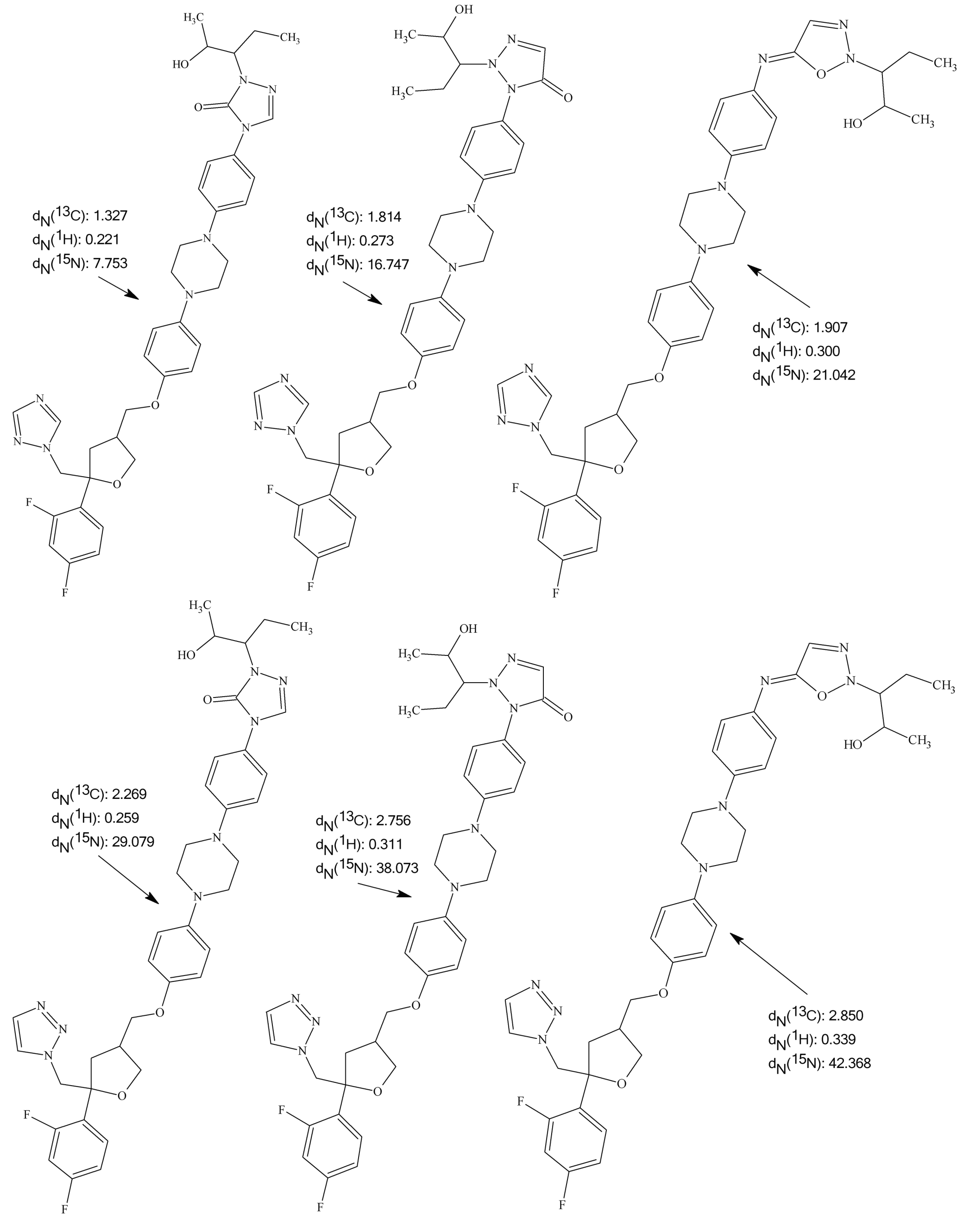


Figure 2. The best six unique structures obtained from a computation run using Elucidator v 12.01 and incorporating all nitrogen constraints. Goodness of fit for the proton, carbon, and nitrogen chemical shifts for each structure are shown. The "best" structure, upper left hand corner, is the correct structure.

(see above). It was clear that additional constraints could have reasonably been added to the dataset based on straightforward logic and in a practical application, this would be done. However, the objective here was to test a large molecule with a minimum number of constraints and to let the program do as much of the structure generation calculation as possible.

After a consistency check of the data, the computation run commenced. In 6 h 0 min 29 s, the run concluded. A total of 16,550 structures were assembled, with 87 structures stored after a structure and chemical shift filter was applied. The best three structures (based on their fit to ACD-calculated proton, carbon, and nitrogen chemical shifts) were the correct structure for posaconazole. Figure 2 shows the first six structures in order of goodness of fit of the NMR data to ACD predictions. At this point, the 20 ^{1}H-^{15}N GHMBCAD experimental constraints were removed, as were six symmetry-generated

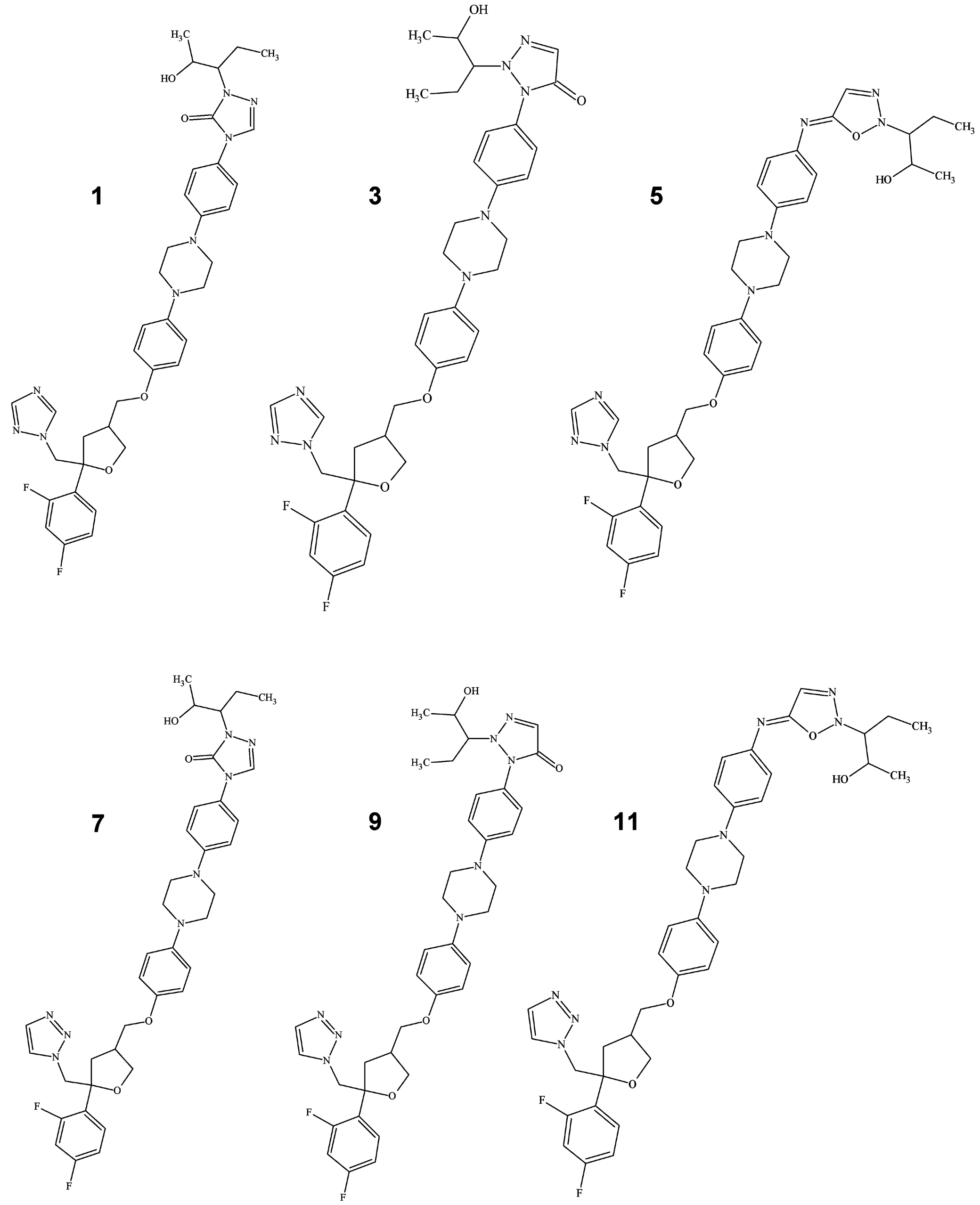


Figure 3. Structure Elucidator run after the addition of 12 GCOSY cross peaks, an adjustment of the intensity threshold for GHMBC cross peaks to 2–3 bond interactions so that the threshold for fourbond interactions was lowered from 3% to 1%, and a change in the filter to prohibit rings from 7 to 30 in size. The run lasted 2 h 50 min; note that these structures are the same as the six most likely structures shown in Figure 2.

constraints directly involving nitrogen atoms, because the initial constraints from interaction between a particular proton and a nitrogen would not exist. The run was restarted. The catastrophic effect of omitting the 26 constraints from the calculation was immediately seen in the number of structures generated; for example, after 25 h, over 104,000,000 structures had been generated, and none had passed the filters and been stored.

The effect of reworking the data was explored briefly in two ways. First, the original data and conditions were reset to match those used for the initial run with nitrogen constraints included. The GCOSY data were then enhanced by marking 12 additional cross peaks in the 2D spectrum. These cross peaks were of lower intensity than those used previously and could be expected to include some four-bond interactions. When the run was restarted, the program ran slightly longer than the initial run at 6 h 16 min, but produced fewer structures, 6193, and fewer structures were saved, 58. Thus, straining to include low-level 2D cross peaks did not enhance the computation time but did reduce the number of candidate structures. The best structure was, again, the correct structure.

A different approach to reworking of the data produced interesting results. Taking the dataset as it stood after the addition of the 12 GCOSY cross peaks, the intensity threshold for GHMBC cross peaks to be used only for 2–3 bond interactions and not for four-bond interactions was lowered from 3 to 1%. At the same time, the structure filter was changed to prohibit rings from 7 to 30 in size—this change was expected to reduce the number of structures saved significantly but not affect the search time. The elucidator run then lasted only 2 h 50 min with 4785 structures generated and only six unique structures were stored. These structures are shown in Figure 3. Again, the best structure was the correct structure.

These latter two examples suggest that adding longer range interactions that are looser in terms of constraints will not improve the structure generation significantly but putting stronger rules for existing constraints can reduce the calculation time dramatically. Thus, constraints that are possibly 4–5 bonds in length are much less useful than short-range interactions. Of course, the last example does have the danger that some cross peaks that represent a four-bond interaction may be misinterpreted. If this is the case, the general result (assuming one is not using “fuzzy” generation [7]) will be an MCD check failure and/or a failure to produce any candidate structures.

A final adjustment of the data was also instructive. All GHMBC long-range constraints were forced to 2–3 bonds. None of these constraints were allowed to extend to four plus bonds. It was anticipated that this action would result in the generation process producing no structures at all. However, 4785 structures were produced (the same as previously) and once again, six unique structures were stored; however, the generation dropped to 59 min. This observation suggests that once the dataset is known or reasonably expected to be self-consistent, that working from a minimalist interpretation of long-range constraints at the outset might prove more efficient. Gradually loosening the constraints and repeating the structure generation step to look for alternative structures could then be explored. It should be noted that using these strong GHMBCAD constraints does not materially affect the elucidation process when the ^{1}H-^{15}N constraints are removed; a repeat of this elucidation was conducted for over 20 h that produced over 13,000,000 structures but, once again, none were stored. The reduction of the number of structures from >104,000,000 to 13,000,000 is a significant reduction but was not sufficient to produce usable structures in a reasonable period of time.

SUMMARY

Posaconazole (**1**) represents a stringent test case for a CASE program because of the presence of eight annular nitrogen atoms each with multiple constraints from the 2D spectra (totaling 20 raw ^{1}H-^{15}N constraints). Köck *et al.* [9] have reported that even partial long-range ^{1}H-^{13}C and ^{1}H-^{15}N heteronuclear shift correlation information can have a dramatic impact on the number of structures that a CASE program generates. The considerable benefit of obtaining ^{15}N 2D data is clear when an investigator is dealing with a nitrogen-containing total unknown is quite clear as shown by this and other published reports [9–13]. We note that this comment can be extended to other types of heteronuclear data, such as ^{19}F as well as ^{31}P data. In fact, as molecules become as large as or even larger than posaconazole, the value of each additional independent constraint in reducing computation time becomes correspondingly larger. A serious problem for CASE studies is how to render certain types of data into a form usable by the CASE program without creating constraints that are not justified by the data. We are continuing our investigations with additional complex molecules and will report on these investigations in the future.

REFERENCES AND NOTES

[1] Martin, G. E.; Williams, A. J. In Encyclopedia of Magnetic Resonance; Harris, R. K.; Wasylishen, R. A., Eds.; Wiley: New York, 2010.

[2] Martin, G. E.; Solntseva, M.; Williams, A. J. In Modern Alkaloids; Fattorusso, E.; Taglialatela-Scafati, O., Eds. Wiley-VCH: New York, 2008; p 409.

[3] Marek, R.; Lyčka, A.; Kolehmainene, E.; Elina, S.; Touse, J. Curr Org Chem 2007, 11, 1154.

[4] Martin, G. E.; Williams, A. J. Annual Report of NMR Spectroscopy, v. 55; Webb, G., Ed.; Academic Press: New York, 2005, pp 1–119.

[5] Marek, R.; Lyčka, A. Curr Org Chem 2002, 6, 35.

[6] Martin, G. E.; Hadden, C. E. J Nat Prod 2000, 63, 543.

[7] Elyashberg, M. E.; Williams, A. J.; Martin, G. E. Prog NMR Spectrosc 2008, 53, 1.

[8] Elyashberg, M.; Blinov, K.; Molodtsov, S.; Smurnyy, Y.; Williams, A. J.; Churanova, T.J Cheminformatics 2009, 1.

[9] Köck, M.; Junker, J.; Lindel, T. Org Lett 1999, 1, 2041.

[10] Nuzillard, J.-M.; Cormolly, J. D.; Delvade, C.; Richard, B.; Zches-Hanrot, M.; Le Me-Olivier, L. Tetrahedron 1999, 55, 11511.

[11] Martin, G. E.; Hadden, C. E.; Russell, D. J.; Kaluzny, B. D.; Guido, J. E.; Duholke, W. K.; Stiemsma, B. A.; Thamann, T. J.; Crouch, R. C.; Blinov, K.; Elyashberg, M. E.; Martirosian, E. R.; Molodtsov, S. G.; Williams, A. J.; Schiff, P. L.; Jr. J Heterocyclic Chem 2002, 39, 1241.

[12] Martin, G. E.; Williams, A. J. ACD/Labs North American User’s Meeting, Princeton, NJ, March 7, 2002.

[13] Grube A. Köck, M. J Nat Prod 2006, 69, 1212.

[14] Cheatham, S.; Kline, M.; Sasaki, R.; Blino, K.; Elyashberg, M.; Molodtsov, S. Magn Reson Chem 2010, 48, 571.

[15] Hilton, B. D.; Feng, W.; Martin, G. E. J Heterocyclic Chem, 2011, 49, in press.

[16] The Structure Elucidator program converts GHMBC correlation information to carbon-carbon bond equivalents; hence, the reference to 1–3 and 1–4 bonds in the text corresponds to the normal 2–4 or 2–5 bond correlations in spectroscopic discussion terms.